Mohamed Khider University of Biskra
Faculty of Exact Sciences
Department of Computer Science

MASTER THESIS

Speciality : Master - Artificial Intelligence - Ref : P1602/2026

TITLE

Explainable AI for Multi-Modality Deep Learning Models in Healthcare

Prepared and defended by

Abdelhalim YOUCEF

Ahmed Madani ARIF

Defense date : 07/06/2026

DEFENSE COMMITTEE

Full Name	Grade	Role
Samir TIGANE	MCA	**President**
Rachida BENABDELAZIZ	Prof	**Supervisor**
Adel Abdelli	MCB	**Examiner**
Nadjoua HABBA	Prof	**Incubator Representative**

Academic Year : 2025/2026

Acknowledgment

First and foremost, we thank Allah for granting us the strength, patience, and perseverance to complete this work.

We would like to express our sincere gratitude to our supervisors, Prof. Rachida Saouli (Benabdelaziz), Dr. Imane Youkana, and Prof. Laid Kahloul, for their valuable guidance, insightful advice, continuous encouragement, and unwavering support throughout the development of this thesis. Their expertise, dedication, and constructive feedback have greatly contributed to the quality of this work and have significantly enriched our academic and research experience.

We would also like to sincerely thank the president of the jury and the examiner for the honour they do us in evaluating and appreciating this work.

We are also very grateful to the Smart Computing Laboratory (LINFI) for providing the necessary facilities, resources, and a stimulating research environment that greatly contributed to the successful realization of this project, the support and conditions provided by the laboratory played an essential role in enabling us to carry out our work in the best possible circumstances.

We would also like to sincerely thank our friends and colleagues for their encouragement, support, and valuable discussions throughout this journey, Their presence, motivation, and shared experiences have made this academic path more enjoyable and meaningful.

Our deepest appreciation goes to our parents and families for their patience and constant encouragement throughout our academic journey, Their continuous support has been a constant source of strength and motivation. We are profoundly grateful for everything they have done to help us reach this stage.

Finally, we extend our sincere thanks to everyone who contributed, directly or indirectly, to the completion of this thesis, through academic guidance, technical assistance, or constructive feedback. Every contribution, no matter how small, has been deeply appreciated and has played a meaningful role in the completion of this work.

Abstract

The rapid adoption of deep learning in clinical settings has produced multimodal systems capable of jointly processing medical images and clinical text. To address the opacity of these systems, Explainable AI is highly needed. We focus on applying Post-hoc Attribution techniques to generative Vision-Language Models, which represent a significant advance in automated visual question answering and radiology report generation directly from Volumetric Medical Imaging data like three-dimensional CT and MRI scans. Despite their clinical potential, these models raise critical concerns about transparency and accountability that constitute a major barrier to their deployment in real-world healthcare environments.

The central contribution of our work is a differentiable attribution framework that enables the consistent application of methods like Integrated Gradients alongside feature-level methods such as Layer Grad-CAM. To resolve a fundamental bottleneck in applying gradient-based methods to generative architectures, a Teacher Forcing strategy converts the stochastic autoregressive generation process into a fully differentiable computation graph, with a scalar target defined by cumulative log-likelihood. To rigorously assess the reliability of these explanations, a Faithfulness Evaluation is introduced.

The framework is evaluated on CT scans from the M3D-Cap and M3D-VQA datasets across three clinical question types. Input-level methods produce sparse and spatially selective attribution maps that shift meaningfully with question type, confirming that the explanations are semantically grounded in the model's clinical reasoning. A voxel deletion protocol establishes the most causally faithful techniques, while other layers exhibit significant spatial diffusion and limited discriminative localization.

A clinical platform RadEx is developed to demonstrate the practical applicability of the proposed framework, allowing clinicians to upload scans, formulate clinical queries in natural language, and receive diagnostic insights alongside visual explanations. The scientific contribution of this work was validated through a peer-reviewed publication accepted at an IEEE conference, as documented in the Scientific production section.

Keywords: Explainable AI, Post-hoc Attribution, Vision-Language Models, Volumetric Medical Imaging, Integrated Gradients, Grad-CAM, Teacher Forcing, Faithfulness Evaluation.

Résumé

L'adoption croissante de l'apprentissage profond dans les environnements cliniques a produit des systèmes multimodaux capables de traiter conjointement des images médicales et du texte clinique. Pour remédier à l'opacité de ces systèmes, l'IA Explicable est devenue essentielle. Nous appliquons les techniques d'Attribution Post-hoc aux architectures génératives telles que les Modèles Vision-Langage, qui représentent une avancée significative permettant la réponse automatisée et la génération de comptes rendus radiologiques directement à partir de données d'Imagerie Médicale Volumétrique telles que les scanners CT et d'IRM. Malgré leur potentiel, ces modèles soulèvent des préoccupations critiques en matière de transparence et de responsabilité qui constituent un obstacle majeur à leur déploiement en conditions réelles.

La contribution centrale de notre travail est un cadre d'attribution différentiable qui permet l'application cohérente de méthodes telles que les Gradients Intégrés, ainsi que de méthodes de caractéristiques comme le Layer Grad-CAM. Pour résoudre un verrou fondamental lié à l'application de méthodes basées sur les gradients aux modèles génératifs, une stratégie de Forçage par l'Enseignant convertit le processus de génération autorégressif stochastique en un graphe de calcul entièrement différentiable, définissant une cible scalaire par la log-vraisemblance cumulée. Afin de vérifier la fiabilité de ces explications, une Évaluation de la Fidélité est rigoureusement introduite.

Le cadre est évalué sur des scanners CT issus des jeux de données M3D-Cap et M3D-VQA pour trois types de questions cliniques. Les méthodes au niveau de l'entrée produisent des cartes d'attribution parcimonieuses et spatialement sélectives qui évoluent de manière significative, confirmant que les explications sont sémantiquement ancrées dans le raisonnement clinique. Un protocole de suppression de voxels établit les techniques les plus causalement fidèles, tandis que d'autres couches présentent une diffusion spatiale significative et une localisation discriminante limitée.

Une plateforme clinique RadEx est développée pour démontrer l'applicabilité pratique du cadre proposé, permettant aux cliniciens de télécharger des examens, de formuler des requêtes en langage naturel et de recevoir des insights diagnostiques accompagnés d'explications visuelles. La contribution scientifique de ce travail a été validée par une publication acceptée dans une conférence IEEE à comité de lecture, comme le documente la section Scientific production.

Mots-clés: IA Explicable, Attribution Post-hoc, Modèles Vision-Langage, Imagerie Médicale Volumétrique, Gradients Intégrés, Grad-CAM, Forçage par l'Enseignant, Évaluation de la Fidélité.

ملخص

أدى التبني السريع للتعلم العميق في البيئات السريرية إلى إنتاج أنظمة متعددة الوسائط قادرة على معالجة الصور الطبية والنصوص السريرية معًا. ولمعالجة غموض هذه الأنظمة، أصبحت الحاجة ماسة إلى **الذكاء الاصطناعي القابل للتفسير (Explainable AI)**. نركز في هذا العمل على تطبيق **تقنيات الإسناد اللاحق (Post-hoc Attribution)** على **النماذج التوليدية للرؤية واللغة (Vision-Language Models)**، والتي تمثل تقدمًا كبيرًا في الإجابة الآلية على الأسئلة المرئية وتوليد التقارير الإشعاعية مباشرة من بيانات التصوير الطبي الحجمي مثل **التصوير المقطعي المحوسب (CT)** و**التصوير بالرنين المغناطيسي (MRI)** ثلاثي الأبعاد. وعلى الرغم من إمكاناتها السريرية، إلا أن هذه النماذج تثير مخاوف حاسمة بشأن الشفافية والمساءلة، مما يشكل حاجزًا رئيسيًا أمام نشرها في بيئات الرعاية الصحية في العالم الحقيقي.

تتمثل المساهمة المركزية لعملنا في تقديم **إطار إسناد قابل للاشتقاق** يتيح التطبيق المتسق لطرق مثل **التدرجات المتكاملة (Integrated Gradients)** إلى جانب الطرق على مستوى السمات مثل **Layer Grad-CAM**. ولحل عقبة أساسية في تطبيق الطرق القائمة على التدرجات على الهياكل التوليدية، تقوم استراتيجية **الإجبار من قبل المعلم (Teacher Forcing)** بتحويل عملية التوليد الانحداري الذاتي العشوائية إلى رسم بياني حسابي قابل للاشتقاق بالكامل، مع هدف قياسي محدد بواسطة احتمالية اللوغاريتم التراكمي. وللتقييم الدقيق لموثوقية هذه التفسيرات، تم إدخال **التقييم الصارم للموثوقية (Faithfulness Evaluation)**.

يتم تقييم هذا الإطار على فحوصات التصوير المقطعي المحوسب من مجموعات بيانات **M3D-Cap** و **M3D-VQA** عبر ثلاثة أنواع من الأسئلة السريرية. تنتج الطرق على مستوى المدخلات (**Input-level methods**) خرائط إسناد متناثرة وانتقائية مكانيًا تتغير بشكل هادف بناءً على نوع السؤال، مما يؤكد أن التفسيرات متجذرة دلاليًا في المنطق السريري للنموذج. ويؤسس **بروتوكول حذف العناصر الحجمية (Voxel Deletion)** التقنيات الأكثر موثوقية من الناحية السببية، في حين تظهر الطبقات الأخرى انتشارًا مكانيًا كبيرًا وتحديدًا تمييزيًا محدودًا.

كما تم تطوير منصة سريرية تدعى **RadEx** لإثبات قابلية التطبيق العملي للإطار المقترح، مما يسمح للأطباء بتحميل الفحوصات، وصياغة الاستفسارات السريرية باللغة الطبيعية، وتلقي الرؤى التشخيصية جنبًا إلى جنب مع التفسيرات المرئية. وقد تم التحقق من المساهمة العلمية لهذا العمل من خلال **ورقة بحثية خضعت لمراجعة الأقران وتم قبولها في مؤتمر IEEE**، كما هو موثق في قسم الإنتاج العلمي.

الكلمات المفتاحية: الذكاء الاصطناعي القابل للتفسير (Explainable AI)، الإسناد اللاحق، نماذج الرؤية واللغة، التصوير الطبي الحجمي، التدرجات المتكاملة، Grad-CAM، إجبار المعلم (Teacher Forcing)، تقييم الموثوقية (Faithfulness Evaluation).

Contents

Abstract **2**

Résumé **3**

List of Figures **12**

1 Artificial Intelligence in Healthcare **16**

1.1 Introduction 16

1.2 Medical Data 16

1.2.1 Medical Imaging 16

1.2.2 Clinical Text 17

1.2.3 Challenges in Medical Data 18

1.3 Deep Learning in Medical Imaging 18

1.3.1 Image Classification Architectures 19

1.3.2 Image Segmentation Architectures 21

1.3.3 Object Detection and Localization Architectures 22

1.3.4 Medical Report Generation Architectures 23

1.4 Multimodal Learning and Vision-Language Models 26

1.4.1 Fusion Strategies 26

1.4.2 Vision-Language Models 27

1.4.3 Medical Vision-Language Models 28

1.5 Conclusion 29

2 Explainable AI in Healthcare **31**

2.1 Introduction . . . 31

2.2 Foundations and Taxonomy . . . 31

2.2.1 Core Concepts and Definitions . . . 32

2.2.2 The Accuracy-Interpretability Trade-off . . . 32

2.2.3 Taxonomy of XAI Methods . . . 33

2.3 Post-hoc Explanation Techniques . . . 37

2.3.1 Gradient-Based Attribution Methods . . . 37

2.3.2 Perturbation-Based Methods . . . 39

2.4 Evaluating Explanation Quality . . . 40

2.4.1 Evaluating Explanations in Generative VLMs . . . 42

2.5 Applications and Challenges . . . 42

2.6 Related Works . . . 45

2.6.1 XAI for Unimodal Medical Image Classification . . . 46

2.6.2 XAI for Medical Vision-Language Models . . . 46

2.6.3 XAI for Generative and Volumetric Settings . . . 47

2.7 Discussion: Limitations of Existing Work . . . 47

2.7.1 Dominance of 2D and Unimodal Settings . . . 48

2.7.2 Focus on Classification Over Generation . . . 48

2.7.3 Lack of Faithfulness Evaluation in Medical VLMs . . . 49

2.7.4 The Gap in Volumetric Multimodal XAI . . . 49

2.8 Motivation and Problem Statement . . . 50

2.8.1 Scientific Motivation . . . 50

2.8.2 Clinical Motivation . . . 51

2.8.3 Problem Statement . . . 51

2.9 Conclusion 52

3 Conception of the System **54**

3.1 Introduction 54

3.2 Work Objectives 54

3.3 System Architecture 55

3.3.1 Data Layer 56

3.3.2 AI Model Layer 56

3.3.3 End-User Layer 56

3.4 Functional Architecture 57

3.4.1 Presentation Of The Model 58

3.4.2 The XAI Module 61

3.4.3 Attribution Map Generation 67

3.5 Evaluation Protocol 68

3.5.1 Voxel Deletion Protocol 68

3.5.2 Sample Selection Criteria 69

3.6 End-User Layer Analysis and Design 70

3.6.1 Functional Requirements (Use Case Diagrams) 70

3.6.2 Static Structure (Class Diagrams) 71

3.6.3 Dynamic Behavior (Sequence Diagrams) 72

3.7 Conclusion 74

4 Implementation and Results **76**

4.1 Introduction 76

4.2 Experimental Setup 76

4.2.1 Dataset 76

4.2.2 Software Framework 77
4.2.3 Hardware Environment 77
4.2.4 Method Configuration 77
4.2.5 Visualization Strategy 78
4.3 Qualitative Results 79
4.3.1 Input-Space Methods Across Question Types 79
4.3.2 Layer Grad-CAM Analysis 83
4.4 Quantitative Faithfulness Evaluation 85
4.4.1 Voxel Deletion Results 85
4.5 Discussion 86
4.5.1 Input-Level Methods Outperform Feature-Level Methods 86
4.5.2 Layer Grad-CAM's Limitations in Multimodal Architectures 87
4.5.3 Question Sensitivity as Evidence of Meaningful Attribution 87
4.5.4 The Faithfulness-Resolution Trade-off 87
4.6 Conclusion 88

5 Platform Implementation 89

5.1 Introduction 89
5.2 Platform Overview 90
5.3 Platform Architecture 91
5.3.1 Frontend Layer 91
5.3.2 API Layer 92
5.3.3 Inference Layer: Volumetric Data Pipeline 92
5.4 Clinical Workflow 93
5.5 Explainability Integration 94

5.5.1 The Explainability Viewer 95
5.5.2 Attribution Maps in Structured Reports 95
5.6 Dashboard Interface and Navigation 96
5.6.1 Navigation Structure 96
5.6.2 Overview 96
5.6.3 Reports 96
5.6.4 Analyses 98
5.7 Academic Platform: RadEx Edu 98
5.7.1 Resident Progress and Milestones 98
5.7.2 Case Modules and Library 99
5.7.3 Interactive Reporting Practice 101
5.8 Conclusion 102

General Conclusion **103**

Scientific production **105**

List of Figures

1.1 Transition from 2D Pixel-based to 3D Voxel-based Imaging. 17

1.2 Operational comparison between 2D and 3D convolution kernels. 20

1.3 Structural evolution from standard sequential CNNs to ResNet and DenseNet. . 20

1.4 The U-Net architecture. 22

1.5 Unrolled computational graph of a standard Recurrent Neural Network. 24

1.6 Internal gating mechanics and dual-stream memory paths of an LSTM unit. . . 25

1.7 The foundational Transformer Encoder-Decoder Architecture [1]. 26

1.8 Conceptual Illustration of Multimodal Fusion Strategies. 27

1.9 General Architecture of a Vision-Language Model. 28

1.10 The MedCLIP architecture for medical image-text contrastive learning. 29

2.1 The Accuracy-Interpretability Trade-off and the Future Research Arena of XAI. 33

2.2 Taxonomy of XAI Methods 36

2.3 Visualizing Faithfulness and Plausibility in Medical XAI 41

3.1 Global system architecture. 55

3.2 Functional computational pipeline of the explainable vision-language system. . . 57

3.3 Architecture and Training Phases of Med3DVLM 59

3.4 Architecture of Multi-Scale MLP-Mixer Projector 61

3.5 Execution Flow of the Dual-Pass XAI Module. 62

3.6 Two-Pass Attribution Framework Architecture 64

3.7 Clinical Use Case 70

3.8 Educational Use Case 70
3.9 Clinical Class Diagram 71
3.10 Educational Class Diagram 72
3.11 Report Generation Sequence 73
3.12 List Reports Sequence 73
3.13 Educational Practice Sequence 74

4.1 Attribution maps across VQA and report generation formats 79
4.2 Closed-ended attribution maps for pelvic CT case 80
4.3 Open-ended attribution maps for pelvic CT case 80
4.4 Report generation attribution maps for pelvic CT case 80
4.5 Closed-ended attribution maps for abdominal CT case 81
4.6 Open-ended attribution maps for abdominal CT case 81
4.7 Report generation attribution maps for abdominal CT case 81
4.8 Closed-ended attribution maps for craniofacial CT case 82
4.9 Open-ended attribution maps for craniofacial CT case 82
4.10 Report generation attribution maps for craniofacial CT case 82
4.11 Comparison of Layer Grad-CAM at consensus vs. max-activation slices 83
4.12 Voxel deletion curves for all attribution methods 85

5.1 RadEx landing page 90
5.2 RadEx volumetric analysis demo 91
5.3 RadEx upload interface 94
5.4 RadEx Overview dashboard 97
5.5 RadEx structured report 97
5.6 RadEx Analyses dashboard 98

5.7 RadEx Edu resident progress tracking dashboard. 99

5.8 RadEx Edu training modules and assigned practice cases. 100

5.9 RadEx Edu searchable clinical case archive. 100

5.10 RadEx Edu diagnostic image viewer and reporting interface. 101

5.11 RadEx AI ground truth comparison for resident feedback. 102

General Introduction

Artificial intelligence (AI) has become a transformative force across numerous domains, with particularly significant impact in healthcare. Advances in deep learning (DL) have enabled the development of highly accurate models for tasks such as medical image analysis, disease diagnosis, prognosis prediction, and clinical decision support. By leveraging large-scale datasets and high-capacity neural architectures, these systems are capable of learning complex representations from heterogeneous data, often achieving state-of-the-art performance across a wide range of applications [2].

In modern clinical practice, medical data are inherently multimodal. Physicians routinely rely on both visual information, such as computed tomography (CT) and magnetic resonance imaging (MRI), and textual data, including radiology reports and electronic health records. To effectively exploit these complementary sources of information, multimodal deep learning has emerged as a powerful paradigm. In particular, vision-language models (VLMs) have demonstrated strong capabilities in jointly processing images and text, enabling applications such as medical visual question answering, report generation, and cross-modal retrieval [3].

Despite their remarkable performance, deep learning models especially large multimodal architectures remain fundamentally opaque. Their decision-making processes are often difficult to interpret, raising critical concerns regarding transparency, reliability, and trust. This issue is particularly crucial in the medical domain, where decisions directly impact patient outcomes and must be supported by clear and clinically meaningful explanations. The lack of interpretability in such systems represents a major barrier to their widespread adoption in real-world healthcare environments.

Explainable Artificial Intelligence (XAI) has been proposed as a solution to this challenge, aiming to make machine learning models more transparent and understandable. Existing XAI approaches can be broadly categorized into intrinsic methods, which incorporate interpretability directly into the model design, and post-hoc methods, which analyze trained models without modifying their internal structure. While intrinsic approaches offer built-in interpretability, they often involve a trade-off with model performance. In contrast, post-hoc methods particularly gradient-based attribution techniques such as Saliency maps, Integrated Gradients, and Grad-CAM provide a flexible and effective way to explain complex deep learning models without compromising predictive accuracy [4]. Although XAI has been extensively studied in the context of unimodal models, its application to multimodal and generative architectures remains relatively underexplored. This is especially true for recent volumetric vision-language models designed for medical imaging, which process high-dimensional 3D data alongside textual inputs. The complexity of these models introduces additional challenges for explainability, including cross-modal interactions, high computational cost, and the stochastic nature of autoregressive generation.

In this context, this work focuses on the study and evaluation of explainability methods for multimodal deep learning models in healthcare, with a particular emphasis on volumetric medical vision-language models. Building upon recent advances in post-hoc attribution techniques, the objective is to investigate how different explainability methods behave when applied to generative multimodal architectures and to assess their ability to produce faithful and spatially meaningful explanations.

To achieve this goal, several gradient-based attribution methods are analyzed and compared within a unified framework. A key contribution of this work lies in adapting these methods to autoregressive generative models by introducing a differentiable attribution objective, enabling consistent and reproducible explanations. The proposed approach is evaluated on a state-of-the-art volumetric medical vision-language model, and its effectiveness is assessed through both qualitative visualization and quantitative faithfulness metrics.

In addition to the research contributions, this work is supported by the development of a practical platform designed as a specialized AI interface for medical analysis. The platform enables users to upload various medical imaging modalities, including CT, MRI, and X-ray scans, and interact with the system through natural language queries. It provides AI-generated diagnostic insights alongside visual explanations within an integrated visualization environment. This implementation serves as a real-world validation of the proposed methods, demonstrating their applicability in clinically oriented scenarios and bridging the gap between theoretical research and practical deployment.

The main contributions of this work can be summarized as follows:

- A comprehensive study of explainable artificial intelligence techniques in the context of multimodal healthcare data.
- The adaptation of post-hoc gradient-based attribution methods to generative volumetric vision-language models.
- The introduction of a differentiable attribution framework enabling consistent explainability in autoregressive settings.
- A comparative evaluation of multiple attribution methods in terms of spatial precision and faithfulness.
- The development of an interactive platform demonstrating the practical applicability of explainable AI in medical imaging.

The remainder of this dissertation is organized as follows. Chapter 1 provides the theoretical foundations of artificial intelligence in healthcare, covering core deep learning paradigms, their application to medical imaging tasks, and the state of the art in multimodal vision-language

models. Chapter 2 establishes the theoretical framework of explainable artificial intelligence, presenting a rigorous taxonomy of XAI methods, reviewing post-hoc gradient-based attribution techniques, and critically examining the limitations of existing work before formally stating the research problem addressed by this work. Chapter 3 presents the methodological framework of the proposed approach, detailing the global system architecture across its distinct layers and presenting the Med3DVLM model. Within this chapter, the core methodological contribution is introduced: the two-pass XAI module based on teacher forcing, followed by the volumetric attribution methods, the quantitative evaluation protocol, and the design analysis of the end-user layer. Chapter 4 covers the implementation and experimental evaluation, reporting qualitative and quantitative results across three clinical question types and discussing their scientific and clinical implications. Chapter 5 presents RadEx AI, the clinical platform developed to demonstrate the practical applicability of the proposed framework in a real-world medical imaging context. Finally, the dissertation concludes with a summary of contributions and outlines perspectives for future research.

Chapter 1

Artificial Intelligence in Healthcare

1.1 Introduction

This chapter establishes the technical and conceptual foundations of artificial intelligence in healthcare by examining the field through an architecture-driven application framework. It begins by exploring the core modalities of medical data, including both volumetric imaging and clinical text, which form the basis for modern diagnostic systems. Subsequently, it reviews the primary deep learning architectures that have driven the most significant advances in medical image analysis. Foundational two-dimensional and volumetric three-dimensional convolutional layers, sequential networks, and global self-attention Transformer mechanisms are analyzed directly within the context of core medical tasks such as image classification, segmentation, object detection, and automated report generation.

The paradigm of multimodal learning is then introduced, with particular attention given to the fusion strategies that enable models to jointly process visual and textual information. Finally, Vision-Language Models (VLMs) are presented as the state of the art in this space, with specific emphasis on medical VLMs such as MedCLIP and Med3DVLM, which represent the most advanced systems currently available for clinical image-text understanding.

1.2 Medical Data

1.2.1 Medical Imaging

Medical imaging constitutes one of the most important sources of information in modern healthcare. Techniques such as computed tomography (CT) and magnetic resonance imaging (MRI) provide detailed visual representations of internal anatomical structures, enabling clinicians to detect abnormalities, monitor disease progression, and guide therapeutic interventions.

Unlike natural images, medical imaging data are often volumetric, represented as three-dimensional (3D) volumes rather than two-dimensional (2D) images. This structural distinction

is illustrated in **Figure 1.1**, which shows the transition from 2D pixel-based grids to 3D voxel-based volumetric data. While a standard image is composed of pixels on a single plane, a typical CT or MRI scan consists of a sequence of slices that form a high-resolution 3D representation. By utilizing voxels, the data preserves the spatial continuity and depth essential for accurate clinical diagnosis, making these complex anatomical structures particularly suitable for deep learning-based analysis [5].

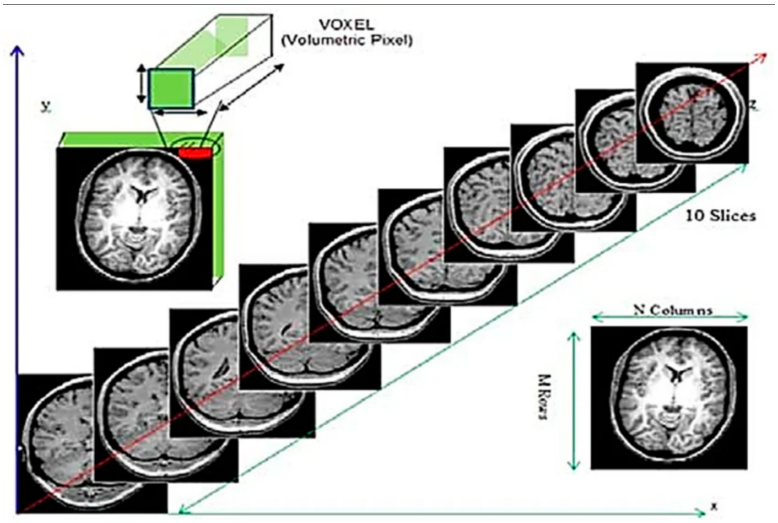


Figure 1.1: Transition from 2D Pixel-based to 3D Voxel-based Imaging. [6]

1.2.2 Clinical Text

In addition to imaging data, clinical text represents a critical component of medical information. This includes radiology reports, physician notes, and electronic health records (EHRs), which contain rich semantic descriptions of patient conditions, diagnoses, and treatments.

Radiology reports, in particular, provide structured interpretations of medical images, linking visual findings to clinical conclusions. These reports often contain domain-specific terminology and complex linguistic structures, making their automatic analysis a challenging task. Natural language processing (NLP) techniques have been increasingly applied to extract meaningful information from clinical text, enabling tasks such as information retrieval, report generation, and clinical decision support [7].

The combination of clinical text with imaging data is especially valuable, as it allows models to learn correspondences between visual features and their semantic interpretations. This multimodal integration is at the core of vision-language models applied to healthcare.

1.2.3 Challenges in Medical Data

Despite the availability of rich medical data, several challenges hinder the effective deployment of AI systems in healthcare.

First, medical data are highly heterogeneous, combining structured data, unstructured text, and high-dimensional imaging modalities. Integrating these diverse sources into a unified learning framework remains a complex task, particularly when aligning information across modalities.

Second, the high dimensionality of volumetric medical images significantly increases computational complexity. Processing 3D data requires specialized architectures and efficient memory management strategies, especially when dealing with large-scale datasets [5].

Finally, deep learning models applied to medical data often suffer from a lack of interpretability. Their "black-box" nature makes it difficult to understand how predictions are generated, which raises concerns about trust, reliability, and clinical adoption. This issue is particularly critical in healthcare, where decisions must be transparent and justifiable [4].

Addressing these challenges requires the development of models that not only achieve high performance but also provide meaningful and interpretable insights, motivating the integration of explainable artificial intelligence techniques in multimodal medical systems.

1.3 Deep Learning in Medical Imaging

AI in medical imaging refers to the use of advanced computational techniques to interpret and analyze medical images with minimal human intervention. It encompasses various methods, including Machine Learning (ML) and Deep Learning (DL), to process large datasets, identify patterns, and make predictions. The scope of AI in this domain includes enhancing image quality, automating analysis, and aiding in complex treatment planning.

Medical imaging encompasses a broad range of modalities, each of which captures different types of anatomical or functional information about the patient. Computed tomography (CT) produces cross-sectional X-ray images that are assembled into three-dimensional volumes, providing detailed views of bones, organs, and soft tissues. Magnetic resonance imaging (MRI) uses strong magnetic fields and radio waves to generate high-contrast images of soft tissues, and is widely used in neuroimaging, musculoskeletal imaging, and cardiac imaging. Conventional radiography (X-ray) remains the most commonly performed imaging examination, used for detecting fractures, pneumonia, and other conditions. Histopathology involves the microscopic examination of stained tissue samples, producing gigapixel-scale whole slide images (WSIs)

that are used for cancer diagnosis and grading. Other modalities include ultrasound, positron emission tomography (PET), retinal fundoscopy, and dermoscopy, each with its own set of characteristics and challenges.

To effectively process these diverse inputs, deep learning research has evolved beyond uniform model configurations. The field has shifted toward task-specific neural architectures engineered to handle unique spatial dimensions, information routing channels, and multimodal alignment goals. The following subsections review the foundational architectures driving the most prominent application areas in medical image analysis.

1.3.1 Image Classification Architectures

Image classification is the fundamental task of assigning a categorical label to an entire input image based on its visual content. In the medical domain, this typically involves identifying the presence of a specific pathology or grading the severity of a disease from scans such as chest X-rays, retinal photographs, or MRI volumes. While early approaches relied on manual feature extraction, the advent of Deep Learning allowed for end-to-end representation learning, where models automatically discover the hierarchical features necessary for diagnostic decision-making.

Foundational CNN Dimension Variants The structural backbone of image classification relies on Convolutional Neural Networks, which must adapt to the geometric constraints of different target imaging modalities. The operational difference between these dimensional approaches is demonstrated in Figure 1.2.

- **2D Convolutional Networks:** Standard 2D CNN configurations utilize planar spatial kernels optimized for static or single-view modalities, such as standard chest X-rays, mammograms, and retinal photographs. These models compute spatial features exclusively across horizontal and vertical pixel coordinates within an isolated grid plane.
- **3D Volumetric Networks:** For advanced cross-sectional clinical imaging modalities like Computed Tomography (CT), Magnetic Resonance Imaging (MRI), or PET scans, standard convolutional blocks are mathematically extended to operate over three dimensions. By processing depth-axis voxels natively, 3D CNN architectures capture contiguous volumetric information, spatial boundary changes, and structural lesions that span across multiple sequential cross-sectional slices.

Structural Evolution of Classification Models Building upon these core 2D and 3D convolutional foundations, the evolution of Deep Learning has introduced specialized architectures

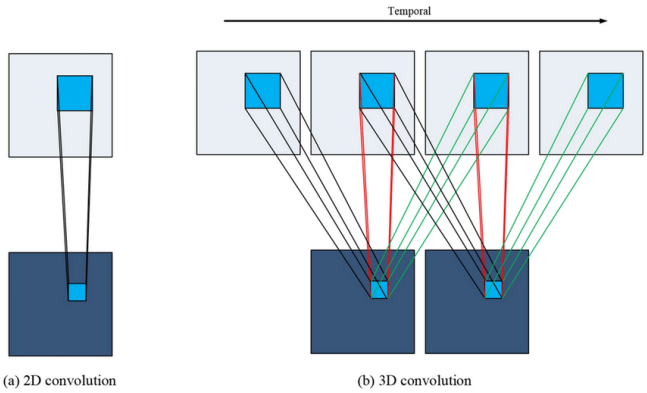


Figure 1.2: Operational comparison between 2D and 3D convolution kernels. [8]

that optimize internal information routing to handle highly complex pathological variations. As illustrated in Figure 1.3, the structural transition from standard CNNs to more advanced models involves a fundamental change in layer connectivity.

While a standard sequential CNN processes features in a strictly linear manner, Residual Networks (ResNet) revolutionized the field by introducing skip connections [9]. As shown in the comparative diagrams of Figure 1.3, these connections allow gradients to bypass one or more layers during backpropagation, effectively solving the vanishing gradient problem and enabling the training of much deeper architectures.

Building on this concept of information flow, Dense Networks (DenseNet) connect each layer to every subsequent layer in a feed-forward fashion [10]. This dense connectivity is clearly visible in the third panel of Figure 1.3, highlighting how the model maximizes feature reuse and reduces total parameters, which is a key advantage for medical datasets where labeled data may be limited. Finally, EfficientNet optimizes this process through a systematic method of compound scaling, which balances network depth, width, and resolution to achieve state-of-the-art accuracy with significantly fewer floating-point operations (FLOPs) [11].

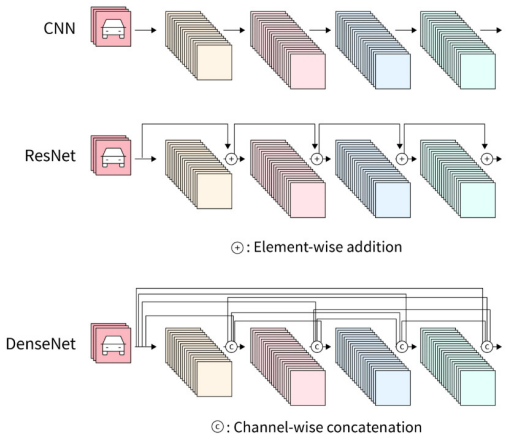


Figure 1.3: Structural evolution from standard sequential CNNs to ResNet and DenseNet.

1.3.2 Image Segmentation Architectures

Image segmentation is the task of assigning a class label to every individual pixel (or voxel, in the case of 3D volumetric data) in an image, thereby delineating the boundaries of anatomical structures or pathological regions. This dense, pixel-level prediction is critical for many clinical applications, including the precise measurement of tumor volume, the planning of intricate surgical interventions, and the automated quantification of organ morphology.

The U-Net Paradigm The U-Net architecture [12] has become the de facto standard framework for medical image segmentation. As illustrated by its distinct symmetric layout in **Figure 1.4**, the network consists of two primary operational paths that form a characteristic "U" shape:

- **The Contracting/Encoder Path (Left Side):** This path acts as a standard convolutional network that repeatedly applies convolutions and pooling operations to capture the abstract semantic context and identify *what* pathologies or structures are present in the image, though it progressively sacrifices spatial resolution.
- **The Expanding/Decoder Path (Right Side):** This path applies up-convolutions to systematically restore the spatial dimensions of the feature maps, allowing the network to precisely locate *where* the boundaries reside.

The Role of Skip Connections The defining architectural innovation of U-Net is its use of horizontal skip connections, visible in **Figure 1.4** as the direct pathways linking the encoder stages across to the decoder stages. These connections crop and concatenate the high-resolution, fine-grained spatial feature maps directly from the contracting path onto the corresponding layers in the expanding path. This mechanism enables the decoding layers to seamlessly combine high-level, abstract semantic context with precise, low-level localized spatial detail, which is essential for preserving sharp boundary definitions around minute clinical lesions.

Architectural Variants Owing to its robust performance, numerous advanced variants of U-Net have been proposed to adapt to specialized medical imaging modalities and tasks:

- **3D U-Net:** Replaces standard 2D operations with 3D convolutions and volumetric max-pooling to process contextual native slices seamlessly across volumetric modalities like MRI and CT.
- **Attention U-Net:** Integrates localized attention gates along the skip connections to suppress irrelevant background regions and actively highlight salient feature responses before concatenation.

- **nnU-Net (no-new-Net):** Introduces a powerful, self-configuring framework that removes manual heuristic tuning by automatically adapting the exact network topology, preprocessing, and training hyper-parameters based on the unique data signature of a given clinical dataset.

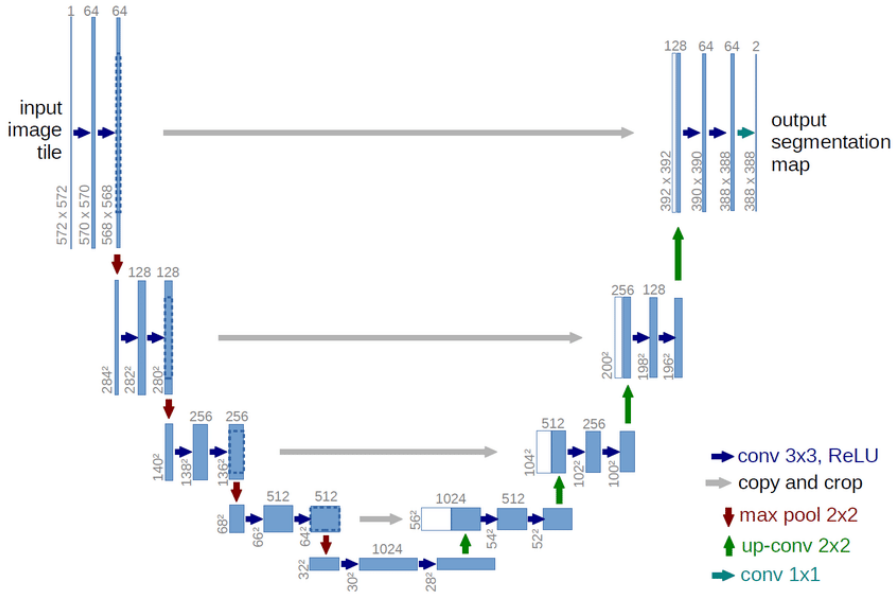


Figure 1.4: The U-Net architecture.
[12]

1.3.3 Object Detection and Localization Architectures

Object detection involves identifying and localizing specific structures or abnormalities within a medical image by predicting bounding boxes around them. This task is particularly relevant for applications such as the detection of pulmonary nodules in CT scans, the localization of lesions in mammograms, and the identification of polyps in colonoscopy images [5].

Faster R-CNN introduced a Region Proposal Network that shares convolutional features with the detection network, enabling end-to-end training and significantly improving both speed and accuracy compared to its predecessors R-CNN and Fast R-CNN [13]. In medical imaging, Faster R-CNN has been widely adopted for lesion detection tasks where precise localization is critical, as its two-stage architecture provides high sensitivity at the cost of moderate inference speed.

YOLO introduced a fundamentally different approach by framing object detection as a single regression problem, dividing the input image into a grid and predicting bounding boxes and class probabilities directly in a single forward pass. This single-stage design makes YOLO significantly faster than two-stage detectors, making it particularly attractive for real-time clinical applications such as endoscopy and ultrasound [14].

RetinaNet addressed the class imbalance problem that typically degrades the performance of single-stage detectors by introducing the Focal Loss function, which down-weights the contribution of easy background examples during training and focuses learning on hard, misclassified foreground objects. This property is especially valuable in medical imaging, where pathological regions are often vastly outnumbered by healthy tissue, making class imbalance a persistent and clinically significant challenge [15].

These architectures have been adapted for medical imaging, often requiring modifications to handle the specific challenges of clinical data, such as small object sizes, low contrast between pathological and healthy tissue, and the need for high sensitivity to avoid missing potentially life-threatening abnormalities

1.3.4 Medical Report Generation Architectures

Automated medical report generation represents a pinnacle task in medical deep learning, shifting the objective from static pixel-level classification or segmentation to dynamic, natural language generation. This task requires an architecture to not only detect subtle pathological features across various clinical modalities, such as lesions in dermoscopic slides, structural anomalies in endoscopic videos, or tissue variations in histopathology, but also to synthesize those complex observations into a structured, free-text diagnostic narrative.

The Encoder-Decoder Mapping Paradigm Because report generation bridges distinct structural domains, it is formulated as an asynchronous image-to-text conversion problem solved via an integrated Encoder-Decoder pipeline:

- **The Visual Encoder:** A deep vision backbone processes the raw medical input data and compresses high-dimensional pixel matrices or volumetric frames into a structured sequence of spatial feature vectors. This backbone is typically implemented as a dense 2D/3D Convolutional Neural Network such as ResNet or DenseNet, which extracts hierarchical spatial features through successive convolutional operations, or as a patch-based Vision Transformer (ViT), which divides the input image into a fixed grid of non-overlapping patches, projects each patch into a token embedding, and processes the resulting sequence through a standard Transformer encoder. The Transformer architecture underlying the ViT design is described in detail in the following section.
- **The Language Decoder:** A sequential text generation network ingests these visual tokens to generate the final clinical narrative token-by-token, conditioned entirely on the extracted visual features.

Evolution of the Language Decoder Generation Network The sequential generation component of this pipeline has evolved across three primary architectural families, transitioning from step-by-step recurrence to fully parallelized global attention mechanisms:

- **Standard Recurrent Neural Networks (RNN):** The foundational sequence processing framework. As illustrated in the unrolled computation graph of Figure 1.5, an RNN processes or generates sequences sequentially by updating an internal hidden state h_t at each time step based on the current input token x_t and the preceding history h_{t-1}. However, repeated weight multiplications over long text fields cause gradients to vanish during Backpropagation Through Time (BPTT), creating a severe long-range information bottleneck [2, 16].

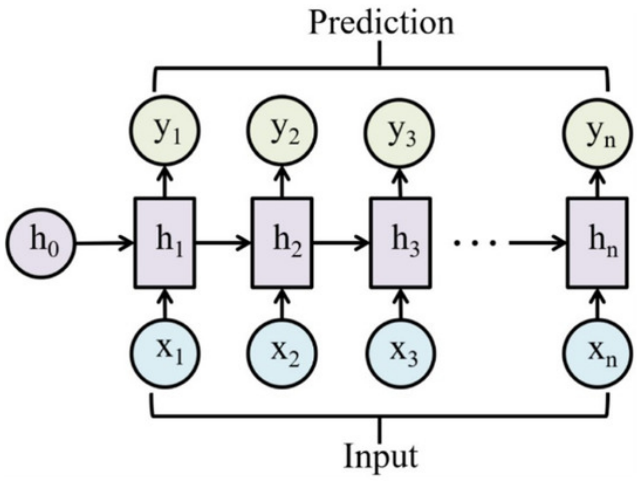


Figure 1.5: Unrolled computational graph of a standard Recurrent Neural Network.

- **Long Short-Term Memory (LSTM):** To resolve the vanishing gradient bottleneck, the LSTM introduces an explicit error-carousel mechanism [16]. As shown in Figure 1.6, it decouples its recurrent routine into a standard hidden path h_t and a dedicated cell state path c_t acting as long-term memory. The information flow along this memory line is tightly regulated by three sigmoidal gating operations: a Forget Gate (f_t) to discard historical data, an Input Gate (i_t) to write new token data, and an Output Gate (o_t) to expose the updated cell state to the next block, ensuring clean linear gradient paths.

- **The Transformer Architecture:** Representing a fundamental paradigm shift, the Transformer completely replaces step-by-step recurrence with a global self-attention mechanism that processes all sequence elements simultaneously. As shown in Figure 1.7, it utilizes a dual-branch layout consisting of an Encoder (left branch) to map token sequences into contextual representations and an auto-regressive Decoder (right branch) to generate the text.

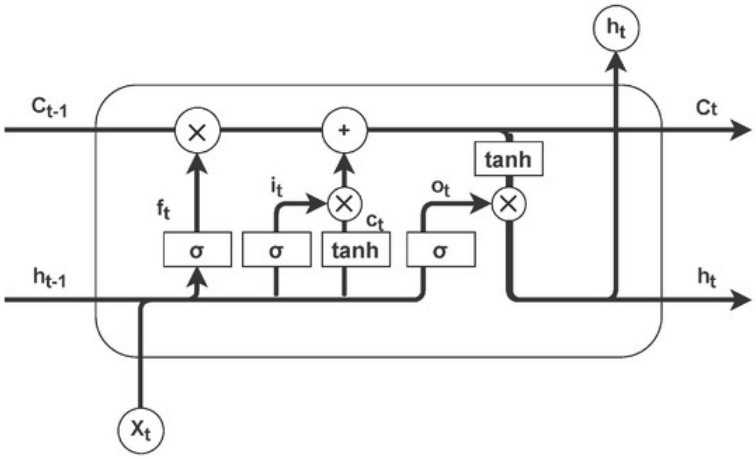


Figure 1.6: Internal gating mechanics and dual-stream memory paths of an LSTM unit.

The core innovation of the Transformer is the scaled dot-product attention mechanism. Given an input sequence, three matrices are computed through learned linear projections: the Query ($\mathbf{Q}$), which represents what the current token is looking for, the Key ($\mathbf{K}$), which represents what each token has to offer, and the Value ($\mathbf{V}$), which contains the actual information to be retrieved. The attention output is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{1.1}$$

where d_k is the key vector dimensionality. The division by $\sqrt{d_k}$ serves as a scaling factor to prevent the dot products from growing too large in magnitude, which would push the softmax function into regions with extremely small gradients.

In practice, the Transformer employs multi-head attention, where multiple attention heads operate in parallel on different linear projections of the input. Each head learns to attend to different aspects of the sequence, allowing the model to simultaneously capture local syntactic relationships, long-range semantic dependencies, and cross-modal correspondences between visual and textual tokens. The outputs of all heads are concatenated and projected back into the model dimension, producing a rich contextual representation of the entire sequence.

Each standard Transformer block combines this parallelized Multi-Head Attention layer with an independent Position-Wise Feedforward Network (FFN), all wrapped inside residual connections and Layer Normalization (LN) links, ensuring stable gradient flow during training.

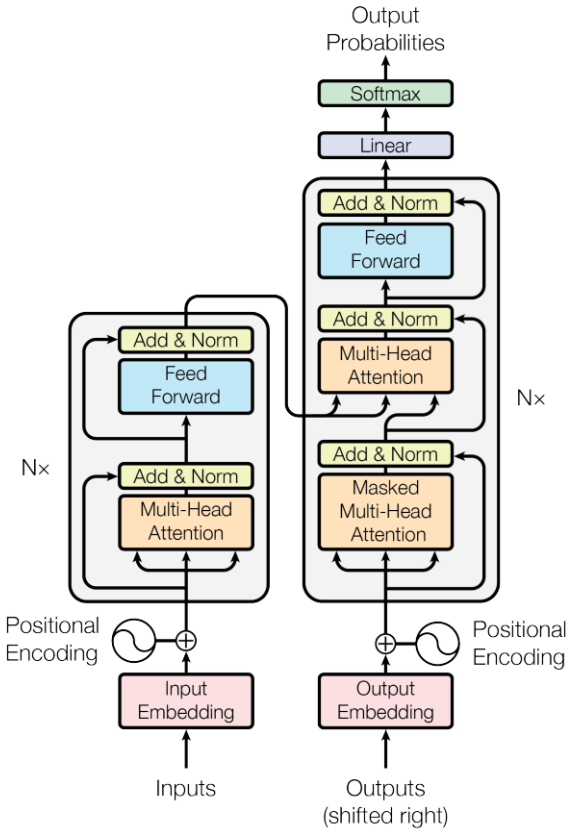


Figure 1.7: The foundational Transformer Encoder-Decoder Architecture [1].

1.4 Multimodal Learning and Vision-Language Models

Multimodal learning refers to the development of models that can process, relate, and integrate information from two or more distinct data modalities, such as images, text, audio, and structured data. The fundamental motivation for multimodal learning is that real-world information is inherently multi-sensory: a clinical diagnosis, for example, draws on visual information from imaging studies, textual information from clinical notes, numerical information from laboratory results, and temporal information from patient history. By learning from multiple modalities jointly, models can capture cross-modal correlations that are invisible to unimodal systems and produce more robust and comprehensive representations of the underlying data.

1.4.1 Fusion Strategies

A central challenge in multimodal learning is the design of the fusion strategy, which determines how and when information from different modalities is combined. The literature broadly categorizes these strategies into three families:early, late, and intermediate fusion, as conceptually illustrated in Figure 1.8

- **Early Fusion (Input-Level Fusion):** As shown in **Figure 1.8a**, raw inputs from different modalities are combined at the earliest stage before being fed into a shared encoder. While this allows the model to learn joint representations immediately, it can be computationally difficult to reconcile modalities with vastly different statistical properties.
- **Late Fusion (Decision-Level Fusion):** In this approach, each modality is processed by an independent encoder, and the results are merged only at the final decision stage (**Figure 1.8b**). While this preserves modality-specific features, it often fails to capture the complex, low-level interactions between data types required for deep clinical reasoning.
- **Middle Fusion (Feature-Level Fusion):** This strategy integrates representations at one or more hidden layers, as depicted in **Figure 1.8c**. Modern architectures often utilize cross-attention mechanisms for this purpose, allowing the model to selectively fuse multiscale features from different modalities into a shared embedding space. This has become the dominant approach in many modern architectures.

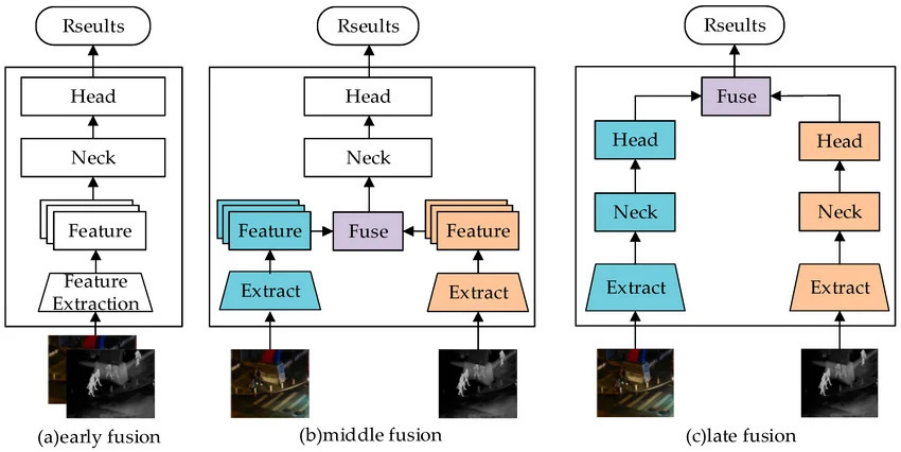


Figure 1.8: Conceptual Illustration of Multimodal Fusion Strategies. [17]

1.4.2 Vision-Language Models

Vision-Language Models (VLMs) are a class of multimodal models specifically designed to jointly process and reason about visual and textual information. The modern era of VLMs was catalyzed by the introduction of CLIP (Contrastive Language-Image Pre-training) [18]. CLIP demonstrated that by training a visual encoder and a text encoder jointly on a very large dataset of image-text pairs using a contrastive learning objective, it was possible to learn a shared embedding space in which semantically related images and texts are mapped to nearby points. This approach enabled zero-shot transfer to a wide variety of downstream tasks without any task-specific fine-tuning, simply by comparing the similarity between image embeddings and text embeddings of candidate labels.

The general architecture of a typical VLM, as illustrated in **Figure 1.9**, consists of three main components. The first component is a visual encoder, which takes an image (or a volume, in the case of 3D data) as input and produces a sequence of visual feature tokens. As depicted in the top branch of the architecture, this encoder is commonly implemented as a CNN, a Vision Transformer (ViT), or a specialized 3D encoder for volumetric data. The second component is a **text encoder** or a language model, which processes the textual input, such as a question, prompt, or caption, and produces a sequence of textual feature tokens.

The third and final component, positioned at the center of the architecture is a fusion or alignment module that combines the visual and textual representations. This can range from simple contrastive alignment (as in CLIP) to more sophisticated cross-attention mechanisms or multimodal projectors that map visual tokens into the embedding space of the language model [18]. This unified design enables the model to bridge the gap between modalities, facilitating complex tasks such as medical report generation and visual question answering.

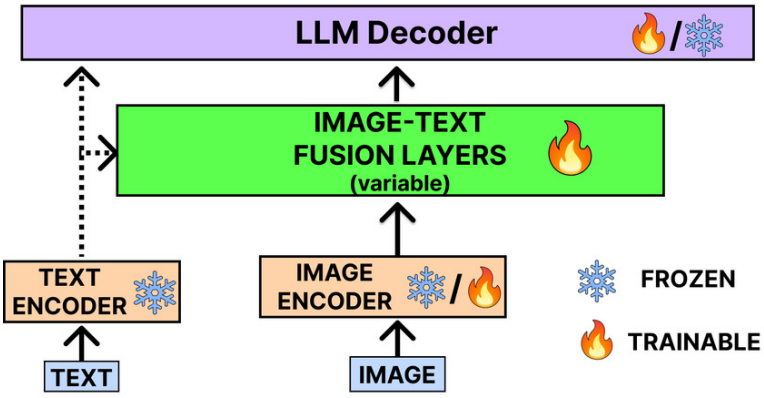


Figure 1.9: General Architecture of a Vision-Language Model.
[19]

1.4.3 Medical Vision-Language Models

The adaptation of VLMs to the medical domain has been an active area of research, driven by the natural correspondence between medical images and their associated clinical reports. While medical VLMs follow the same general architecture as their general-purpose counterparts, they introduce domain-specific adaptations at each component to handle the unique characteristics of clinical data.

As illustrated in Figure 1.10, this specialized structure replaces generic contrastive learning frameworks with a supervision strategy decoupled from rigid image-text pairing constraints. Instead of relying on strict one-to-one paired matching, a Knowledge Extractor pulls fine-grained entities from clinical reports alongside categorical labels from medical images to build a soft target semantic similarity matrix s. A sampled batch of reports and a batch of images are passed through a Text Encoder and an Image Encoder respectively to perform embedding extraction,

yielding feature representations $(v_1, \ldots, v_n)$ and $(t_1, \ldots, t_n)$. These vectors are multiplied to form a predicted similarity matrix $\hat{s}$, which is optimized against the soft semantic targets via a specialized Semantic Matching Loss.

Medical VLMs are designed to perform tasks such as medical visual question answering (VQA), report generation, and clinical reasoning. Early architectures like MedCLIP [20] extended the CLIP contrastive learning framework to paired chest X-rays and reports using this soft semantic alignment, demonstrating improved performance over general-purpose models by leveraging both text and medical images with labels. BioViL [21] and CheXagent [22] further incorporated biomedical text pretraining to enable richer reasoning. More recently, the field has moved toward generative medical VLMs, such as LLaVA-Med [23] and Med3DVLM [24] which utilizes 3D convolutional backbones like DCFormer and employs a middle fusion strategy to capture cross-modal interactions at multiple levels of abstraction while maintaining specialized processing pathways, producing free-text diagnostic responses and performing multimodal reasoning on high-resolution volumetric scans.

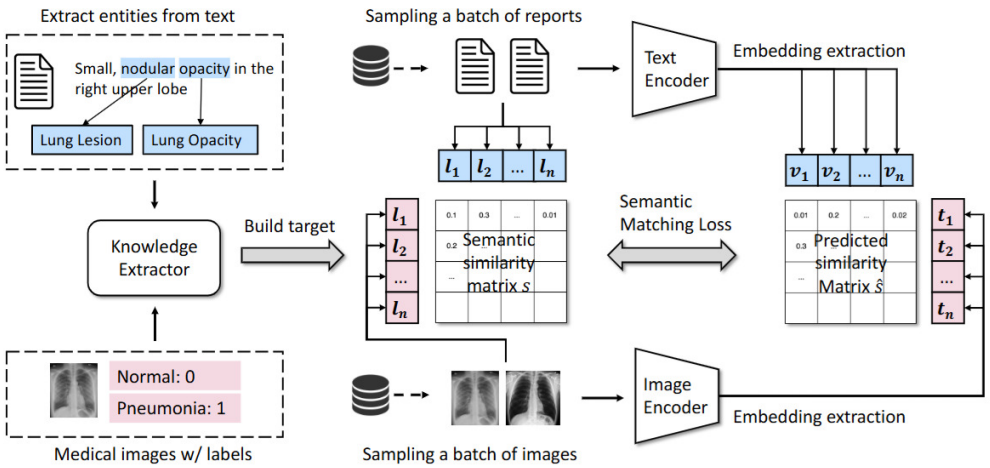


Figure 1.10: The MedCLIP architecture for medical image-text contrastive learning.

1.5 Conclusion

This chapter has established the technical and conceptual foundations of artificial intelligence in healthcare. The core machine learning paradigms were reviewed, followed by an examination of the deep learning architectures that have driven the most significant advances in medical image analysis, from convolutional neural networks to Transformer-based models. The application of these architectures to key medical imaging tasks, including classification, segmentation, object detection, and report generation, was then surveyed, demonstrating the breadth and clinical impact of modern AI systems.

The paradigm of multimodal learning was subsequently introduced, with particular atten-

tion to the fusion strategies that enable models to jointly process visual and textual information. Vision-language models were presented as the state of the art in this space, with medical VLMs such as MedCLIP and Med3DVLM representing the most advanced systems currently available for clinical image-text understanding.

Despite their remarkable performance, these models remain fundamentally opaque. Their decision-making processes are difficult to interpret, raising serious concerns about transparency, trustworthiness, and clinical accountability that represent a critical barrier to their adoption in real-world healthcare environments. Addressing this challenge is the focus of the following chapter.

Chapter 2

Explainable AI in Healthcare

2.1 Introduction

The opacity of modern deep learning models remains a major obstacle to their safe deployment in healthcare, where transparency and accountability are clinical and regulatory requirements. As these systems evolve from conventional convolutional models to multimodal vision-language architectures that jointly process 3D medical volumes and clinical text, it becomes increasingly difficult to understand how a given prediction is produced. In such high-stakes settings, strong predictive performance alone is insufficient if the model cannot justify its conclusions. [4]

Explainable Artificial Intelligence (XAI) has emerged to address this problem by making AI decisions more understandable without sacrificing their clinical utility. In healthcare, XAI helps clinicians interpret and verify predictions while also helping researchers identify model failures before deployment.

This chapter reviews the XAI landscape relevant to this work. It first introduces key concepts and a taxonomy of XAI methods, then presents the post-hoc gradient-based attribution techniques that form the methodological core of this work. It also surveys evaluation frameworks, applications and challenges in healthcare, and the main limitations of current research before defining the scientific and clinical motivation of the work and stating the research problem it addresses.

2.2 Foundations and Taxonomy

The field of Explainable Artificial Intelligence (XAI) is concerned with the development of methods, techniques, and tools that make the decisions and internal processes of AI systems understandable to human users. As discussed in the introduction, the opacity of modern deep learning models has created a critical barrier to their adoption in high-stakes domains such as healthcare. This section establishes the foundational concepts, definitions, and the taxonomy

that structures the XAI landscape [25]

2.2.1 Core Concepts and Definitions

To establish a rigorous framework for the discussion of XAI, it is necessary to carefully define the key concepts that form the foundation of this field. While these terms are often used interchangeably in informal discourse, they represent distinct and complementary aspects of making AI systems understandable [4]:

- **Interpretability** refers to the degree to which a human observer can understand the cause of a specific decision made by a model. An interpretable model allows the user to comprehend how changes in the input would affect the output, and to identify which input features contributed most strongly to a particular prediction.
- **Transparency** refers to the openness and accessibility of a model's internal structure, including its architecture, parameters, and training procedure. A transparent model is one whose entire computational pathway from input to output can be inspected and understood.
- **Explainability** refers to the extent to which the internal mechanics of a machine learning model can be communicated to a human in a meaningful and actionable way. While interpretability focuses on understanding how the model works, explainability goes a step further by focusing on why the model made a particular decision.
- **Fairness** refers to the assurance that AI systems do not produce biased or discriminatory outcomes based on sensitive attributes such as race, gender, age, or socioeconomic status. In the medical domain, fairness is of paramount importance because a model that performs well on one patient demographic but poorly on another could lead to systematic disparities in the quality of care.

These concepts are not isolated but are deeply interconnected. Transparency in a model's structure facilitates interpretability, which in turn enables the generation of meaningful explanations. Explanations can then be used to audit the model for fairness, creating a virtuous cycle that supports the responsible development and deployment of AI systems in healthcare.

2.2.2 The Accuracy-Interpretability Trade-off

A fundamental tension in machine learning is the inverse relationship between predictive performance and model interpretability. This phenomenon is visualized in **Figure 2.1**, which maps various algorithmic families along a Pareto frontier. At the high-interpretability end

of the axis, models such as rule-based learning and linear regression allow for immediate human understanding, yet they often fall short in accuracy when processing complex medical data. Conversely, deep learning and ensemble methods occupy the high-accuracy region but are limited by their "black-box" nature.

As illustrated by the green shaded region in **Figure 2.1**, the "XAI's future research arena" focuses on pushing this frontier outward. This is achieved through two primary pathways. The first involves developing hybrid or interpretability-preserving modeling approaches to raise the accuracy of inherently transparent models. The second, which is the focus of this research, utilizes post-hoc explainability techniques to move complex, high-performance models toward the right, enhancing their interpretability without sacrificing their diagnostic precision. By bridging this gap, we can leverage powerful architectures like Med3DVLM while providing the transparency required for clinical trust.

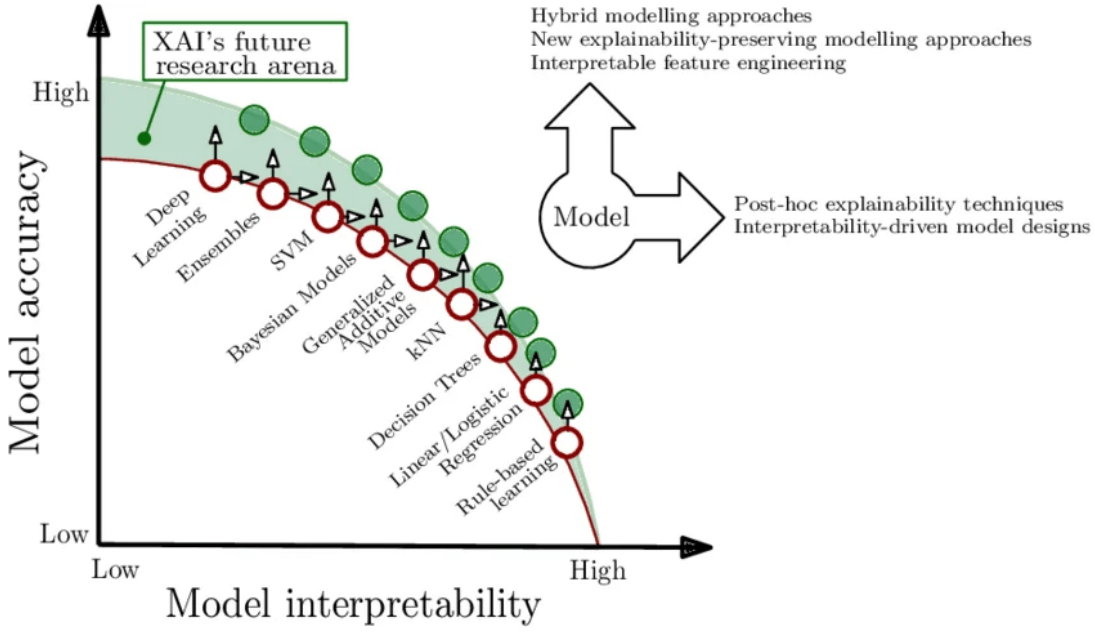


Figure 2.1: The Accuracy-Interpretability Trade-off and the Future Research Arena of XAI.

2.2.3 Taxonomy of XAI Methods

The diverse landscape of XAI techniques can be organized along several orthogonal dimensions, each of which captures a different aspect of how an explanation is generated and what kind of insight it provides [4]. This multi-dimensional classification scheme is systematically summarized in **Figure 2.2**.

2.2.3.1 Intrinsic vs. Post-hoc Methods

The first and perhaps most fundamental dimension of the taxonomy distinguishes methods based on when the explanation is generated relative to the model's training process [25]. As illustrated in the temporal classification branch of **Figure 2.2**, these are divided into two main categories:

Intrinsic (Ante-hoc) Methods These methods produce explanations by virtue of the model's own internal architecture. The model is designed from the outset to be naturally interpretable, meaning that its transparent structural parameters allow a human to understand its decision boundary directly. Standard examples of intrinsically interpretable architectures include linear regression, logistic regression, shallow decision trees, and rule-based systems [26].

Post-hoc Methods Conversely, these techniques generate explanations after a complex model has completed its training phase, without requiring any modifications to the underlying architecture or parameters. Post-hoc methods treat the trained network as a fixed entity and deploy external algorithms to analyze the relationship between inputs and outputs. This approach has the significant advantage of being universally applicable to highly complex architectures, such as deep convolutional or transformer networks, making it a highly flexible branch in the taxonomy.

2.2.3.2 Model-Specific vs. Model-Agnostic Methods

Another core structural division concerns the technical dependency of the explanation algorithm on the underlying machine learning architecture, represented by the model-dependency branch in **Figure 2.2**:

Model-Specific Methods These techniques are engineered to work exclusively with a particular type of model architecture, directly exploiting its internal components (such as weights, activations, or layer configurations) to generate explanations. For instance, Grad-CAM is inherently model-specific to convolutional neural networks because it relies on extracting feature maps and gradient information flowing into a specific convolutional layer to construct a class-discriminative heatmap [27].

Model-Agnostic Methods In contrast, model-agnostic techniques operate independently of the internal network structure. They treat the predictive model as a complete black box, evaluating its behavior solely by perturbing input features and monitoring the corresponding

shifts in output probabilities. Well-known examples within this category include LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) [28, 29].

2.2.3.3 Global vs. Local Explanations

The third dimension defines the operational scope of the explanation, determining whether the interpretive insight covers the entire system or isolates a single inference case, as categorized under the scope branch of **Figure 2.2**:

Global Explanations Global explanation methods aim to describe the holistic behavior of a model across an entire evaluation dataset. Instead of looking at individual instances, they reveal the general rules and feature importances the model relies upon, answering broader macro-level questions such as: "What visual features does the model generally consider most important for diagnosing pneumonia across all patients?"

Local Explanations Conversely, local explanation methods target a single, specific prediction made by the model for a unique input instance. They isolate the exact variables driving an individual output, answering micro-level questions such as: "Why did the model predict that this specific patient has pneumonia based on this chest scan?" In clinical practice, local explanations are often more valuable than global ones because they directly support a clinician's immediate diagnostic decision-making process for individual patient cases.

2.2.3.4 Explanation Output Format

A fourth dimension of the taxonomy concerns the modality and format in which the explanation is communicated to the end user. This dimension is highly relevant in clinical settings, where the effectiveness of an explanation depends on its ability to be rapidly understood and acted upon by a non-technical user such as a physician [30]. As broken down in the output modality branch of **Figure 2.2**, these interfaces span several distinct formats:

Visualization The explanation is presented as a visual artifact, such as a feature attribution map, a saliency map, or a localized bounding region overlaid directly on the input data. This is the most prevalent output format in medical imaging because it mirrors the traditional visual workflows of radiologists, highlighting exactly which spatial regions or voxels triggered the diagnostic flag.

Textual Explanations The explanation is expressed as a natural language text description detailing the model's reasoning steps. This format is uniquely suited to multimodal report generation systems, where the model must justify its internal conclusions using standard clinical terminology.

Surrogate Models This approach fits a secondary, highly interpretable surrogate model (like a shallow linear model) to locally approximate the decision boundaries of a complex black-box model within the immediate neighborhood of a specific data point. LIME is a prominent example of this method, using local surrogacy to achieve proxy faithfulness.

Arguments The explanation is structured as a formal logical argument or a strict set of conditional rules justifying the final output. This format is closely tied to case-based and deductive reasoning engines, providing clean audit trails in high-stakes regulatory environments.

In the context of this work, the primary output format of interest isvisualization, as the proposed XAI platform delivers gradient-based attribution maps overlaid directly onto volumetric medical scans. This format was chosen because it aligns naturally with the visual reasoning process of clinicians and provides an immediately actionable form of explanation in clinical practice.

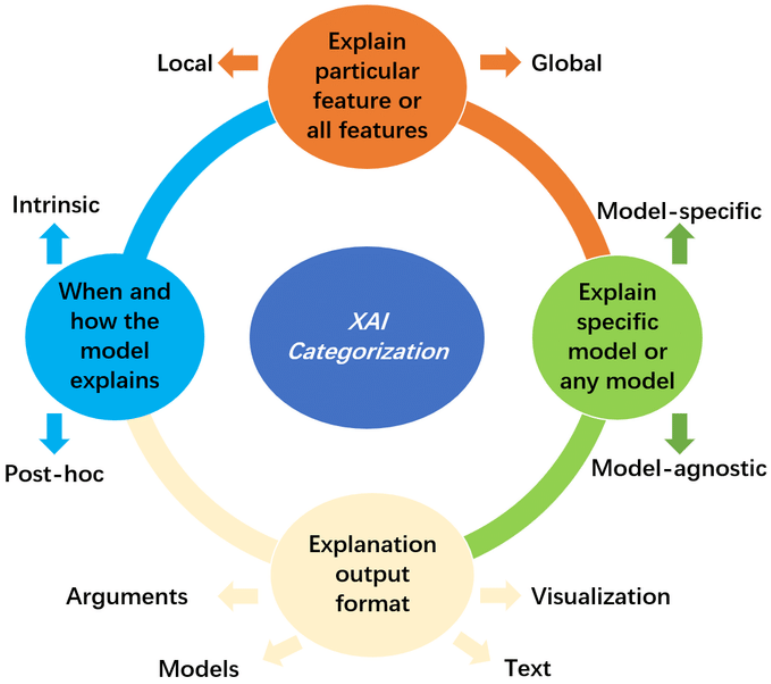


Figure 2.2: Taxonomy of XAI methods categorized by temporal application, architectural dependency, explanation scope, and output format [31].

2.3 Post-hoc Explanation Techniques

This section provides a detailed review of the specific post-hoc explanation techniques that are most relevant to the present dissertation. These methods can be broadly categorized into gradient-based methods, which compute the derivative of the model's output with respect to its input to determine feature importance, and perturbation-based methods, which measure feature importance by observing how the model's output changes when parts of the input are modified or removed.

2.3.1 Gradient-Based Attribution Methods

Gradient-based attribution methods leverage the gradient of the model's output with respect to its input to produce a relevance map that indicates the importance of each input feature. The fundamental idea is that the gradient $\frac{\partial S}{\partial x_i}$ of the model's output S with respect to an input feature x_i measures the sensitivity of the output to small changes in that feature.

2.3.1.1 Input-Level Saliency

Input-Level Saliency, also known as the vanilla gradient method, is the simplest gradient-based attribution technique. It computes the absolute value of the gradient of the model's output with respect to each input element. For an input $\mathbf{x}$ and a scalar output $S(\mathbf{x})$, the saliency map is computed as:

$$R_{\text{Sal}}(x_i) = \left| \frac{\partial S(\mathbf{x})}{\partial x_i} \right| \tag{2.1}$$

The resulting saliency map has the same dimensionality as the input and highlights the input elements where small changes would have the largest effect on the model's output. While simple and computationally efficient, saliency maps are known to suffer from visual noise, gradient saturation, and sensitivity to irrelevant input variations [32].

2.3.1.2 Integrated Gradients

Integrated Gradients (IG) is a principled method for attributing the output of a deep network to its input features. IG addresses the limitations of vanilla saliency by accumulating gradients along a straight-line path from a baseline input $\mathbf{x}_0$ (typically an all-zeros input) to the actual input $\mathbf{x}$:

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_{0,i}) \times \int_0^1 \frac{\partial S(\mathbf{x}_0 + \alpha(\mathbf{x} - \mathbf{x}_0))}{\partial x_i}\, d\alpha \tag{2.2}$$

Integrated Gradients satisfies two important axioms. The completeness axiom guarantees that the sum of all attributions equals the difference between the model's output at the actual input and its output at the baseline: $\sum_i \mathrm{IG}_i(\mathbf{x}) = S(\mathbf{x}) - S(\mathbf{x}_0)$. The sensitivity axiom guarantees that if a feature differs between the input and the baseline and produces a different output, then that feature will receive a non-zero attribution [33].

2.3.1.3 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) belongs to the family of Class Activation Mapping (CAM) methods. it generalized the original CAM idea to any differentiable model by replacing the classification weights with gradient-based importance weights. For a target layer ℓ with feature maps $\mathbf{F}^\ell \in \mathbb{R}^{K \times H' \times W'}$, Grad-CAM first computes importance weights for each channel by global average pooling of the gradient:

$$w_k = \frac{1}{H' \times W'} \sum_h \sum_w \frac{\partial S}{\partial F^\ell_{k,h,w}} \tag{2.3}$$

The Grad-CAM heatmap is then computed as a weighted combination of the feature maps, followed by a ReLU activation:

$$R_{\text{Grad-CAM}} = \mathrm{ReLU}\left(\sum_{k=1}^{K} w_k \cdot \mathbf{F}^\ell_k\right) \tag{2.4}$$

Grad-CAM produces smooth, class-discriminative attribution maps that are easy to interpret visually. However, because it operates at the resolution of the target layer, Grad-CAM heatmaps tend to be spatially diffuse and may fail to capture fine-grained details [27].

2.3.1.4 Guided Backpropagation

Guided Backpropagation is a modified backpropagation procedure that produces high-resolution feature visualizations by suppressing the backward flow of negative gradients through ReLU activations. The guided gradient is computed as:

$$\tilde{\delta} = \mathbf{1}[z > 0] \cdot \mathbf{1}[\delta > 0] \cdot \delta \tag{2.5}$$

This double masking has the effect of propagating only "positive evidence" backward through the network, producing feature visualizations that are sharper and more visually interpretable than standard saliency maps [34].

2.3.1.5 Guided Grad-CAM

Guided Grad-CAM combines the complementary strengths of Grad-CAM and Guided Backpropagation to produce attribution maps that are both class-discriminative and high-resolution. The method works by computing the Grad-CAM heatmap and the Guided Backpropagation map independently, then fusing them through element-wise multiplication:

$$R_{\text{Guided Grad-CAM}} = \text{Norm}\left(R^{\uparrow}_{\text{Grad-CAM}} \odot R_{\text{Guided BP}}\right) \tag{2.6}$$

where $R^{\uparrow}_{\text{Grad-CAM}}$ denotes the Grad-CAM heatmap upsampled to the input resolution and $\odot$ denotes element-wise multiplication.

2.3.1.6 Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) is a backpropagation-based attribution method that redistributes the model's output score backward through the network layer by layer, assigning a relevance score to each neuron according to a set of propagation rules. Unlike gradient-based methods, LRP aims to decompose the output score into contributions from individual input features in a way that satisfies a conservation property: the total relevance is preserved at each layer during backpropagation [35].

2.3.2 Perturbation-Based Methods

Perturbation-based methods assess feature importance by systematically modifying parts of the input and observing the effect on the model's output. Unlike gradient-based methods, perturbation-based methods require multiple forward passes through the model, making them model-agnostic but computationally more expensive.

2.3.2.1 LIME

Local Interpretable Model-agnostic Explanations (LIME) explains individual predictions by learning a simple, interpretable surrogate model in the local neighborhood of the input being. LIME works by generating a set of perturbed versions of the input, obtaining the model's

predictions for each perturbed input, and then fitting a weighted linear model to these input-output pairs [28].

2.3.2.2 SHAP

SHapley Additive exPlanations (SHAP) is a unified framework for feature attribution that is grounded in cooperative game theory. SHAP assigns each feature a Shapley value, which represents the average marginal contribution of that feature to the model's prediction across all possible subsets of features. SHAP values possess several desirable theoretical properties, including local accuracy, consistency, and missingness [29].

2.3.2.3 Occlusion Sensitivity

Occlusion sensitivity is a straightforward perturbation-based method that measures the importance of each region of the input by systematically occluding it and observing the change in the model's output [36]. A sliding window of a fixed size is moved across the input, and at each position, the occluded input is passed through the model.

2.4 Evaluating Explanation Quality

The evaluation of XAI methods is a challenging and actively researched problem, as there is no single, universally accepted metric for measuring the quality of an explanation. To provide a comprehensive assessment, researchers categorize evaluation into a three-tiered hierarchical framework: functional-grounded evaluation (using proxy computational metrics), human-grounded evaluation (using laypeople to test general clarity), and application-grounded evaluation (using domain experts to test real-world utility) [37]. Within these broad testing methodologies, several core performance criteria are utilized to judge the explanation's validity.

Faithfulness (Fidelity) This foundational metric measures how accurately an explanation reflects the true underlying reasoning process of the black-box model. It ensures the explanation is an honest account of the model's math, rather than a superficial artifact. Faithfulness is typically quantified using removability or perturbation tests, where the most influential pixels or voxels highlighted by the attribution map, such as the salient features shown in **Figure 2.3(B)** are systematically masked to monitor the subsequent drop in model confidence [36].

Plausibility Plausibility defines how convincing the explanation appears to a human user and how closely it aligns with established domain knowledge. In medical machine learning, this

is often quantified by treating the explanation as a proxy segmentation mask and calculating the spatial overlap, such as the Intersection over Union (IoU) or Dice score, between the AI's visual attribution map and a clinician's manual ground-truth annotation, as represented by comparing panels **(A)** and **(B)** in **Figure 2.3**.

Stability (Robustness) A reliable explanation must exhibit local stability, meaning that minor, non-semantic variations in the input data should not result in drastically different explanation outputs. In a clinical workflow, for instance, applying slight geometric rotations, translation shifts, or minor acquisition noise to a 3D MRI scan should yield practically identical feature attribution maps, proving the explanation algorithm is tracking true pathology rather than random computational artifacts [38].

Comprehensibility (Interpretability) This criterion characterizes the ease with which a targeted human end-user can process and mentally digest the provided explanation. For medical Vision-Language Models (VLMs), comprehensibility scales beyond pixel tracking to evaluate the semantic clarity, readability, and stylistic formatting of the generated text, ensuring it adheres closely to standard clinical reporting styles and can be rapidly acted upon during a diagnostic emergency.

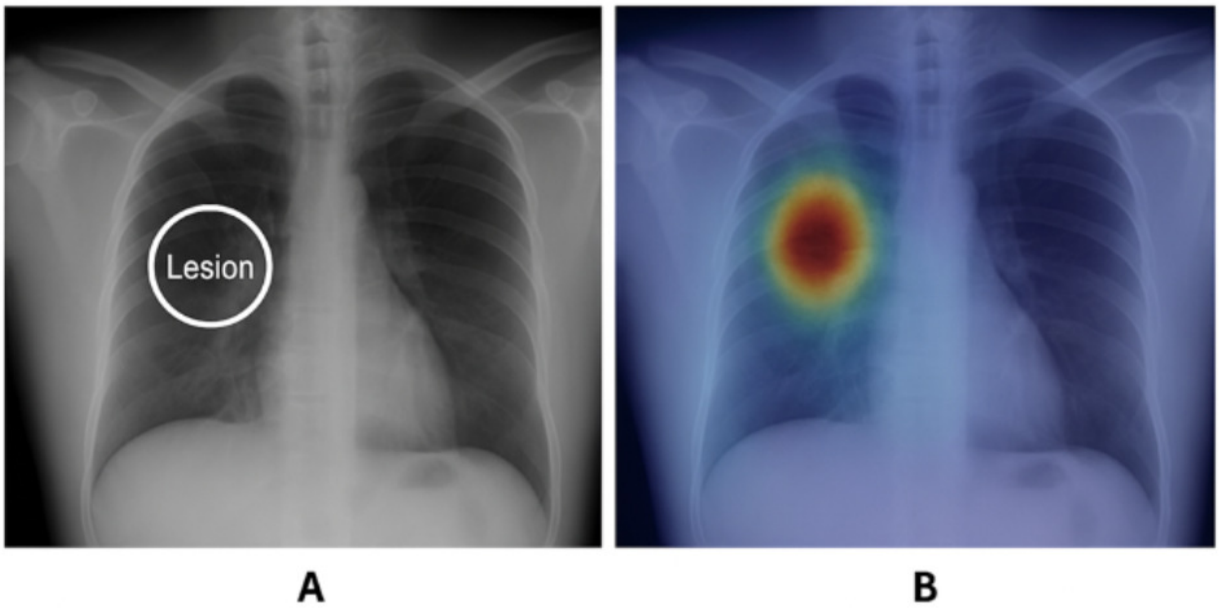


Figure 2.3: Visual comparison between (A) a radiologist-annotated input lesion and (B) an XAI-generated saliency map to evaluate explanation plausibility and faithfulness.

2.4.1 Evaluating Explanations in Generative VLMs

Evaluating post-hoc explanations for generative models requires a paradigm shift from traditional classification-based metrics. Since the model output is a natural language sequence rather than a discrete label, the explanation must demonstrate Multimodal Grounding, ensuring that generated clinical terms correspond to the correct anatomical regions in the 3D volume.

- **Visual-Textual Alignment:** This measures the spatial correspondence between the "attended" visual regions and the specific tokens generated in the radiology report. A faithful explanation for a generative VLM must demonstrate that a specific clinical term, such as "pleural effusion," is directly triggered by the relevant voxel intensities in the 3D scan [39].
- **Generative Faithfulness (Causal Tracing):** Beyond simple feature deletion, faithfulness in generative contexts is assessed by perturbing visual features and observing logical consistency in the output text. For instance, if a detected lesion is masked in the input volume, a faithful model should cease to generate associated pathological descriptors in its report [3].
- **Clinical Factuality and Grounding:** In the medical domain, the quality of an explanation is intrinsically linked to the factuality of the narrative it supports. Metrics such as **RadGraph** are increasingly used alongside XAI attribution maps to ensure that the visual explanation aligns with a clinically accurate and structured textual representation [3].

2.5 Applications and Challenges

The application of XAI methods to healthcare AI systems has received growing attention in recent years, driven by the critical need for transparency, accountability, and trust in medical decision-making. As deep learning models are increasingly deployed in high-stakes clinical environments, the ability to explain and justify their predictions has become not only a scientific priority but also a regulatory and ethical imperative.

XAI methods have been applied across a wide range of medical specialties and clinical tasks, demonstrating their potential to improve diagnostic accuracy, support clinical decision-making, and build trust between AI systems and medical professionals. The following subsections highlight the most prominent application areas

- **Radiology and Medical Imaging:** Gradient-based attribution methods, particularly

Grad-CAM and its variants, have been widely applied to radiology tasks to provide visual explanations for CNN-based diagnostic predictions. In chest X-ray analysis, Grad-CAM heatmaps have been used to highlight the pulmonary regions that most strongly drive a model's pneumonia prediction, enabling radiologists to verify that the model attends to clinically meaningful areas rather than spurious background features. Beyond classification, XAI methods have been applied to segmentation and detection tasks, where attribution maps help clinicians understand which spatial features drive boundary delineation or lesion localization decisions.

- **Pathology and Histological Analysis:** In computational pathology, attention-based multiple instance learning (MIL) models assign importance scores to individual image patches extracted from whole slide images, providing a form of intrinsic interpretability that highlights the most diagnostically relevant tissue regions. These attention maps allow pathologists to visually inspect which microscopic features, such as nuclear pleomorphism or mitotic figures, most strongly influence the model's cancer grading decision. This form of explanation is particularly valuable for rare pathologies where expert consensus is difficult to obtain and model transparency directly supports diagnostic confidence.

- **Clinical Decision Support Systems:** SHAP values have been extensively used in tabular clinical data settings to decompose a model's prediction into individual feature contributions, showing clinicians which patient characteristics, such as age, comorbidities, or laboratory results, are driving a model's risk assessment [29]. In intensive care settings, SHAP-based explanations have been applied to early warning systems for sepsis and organ failure, allowing clinicians to understand why the model flagged a particular patient as high-risk and to cross-check the model's reasoning against their own clinical judgment [30].

- **Medical Visual Question Answering and Report Generation:** In generative multimodal models, XAI methods can attribute the model's generated text responses to specific regions of the input volume, providing clinicians with visual evidence for AI-generated diagnostic conclusions. Gradient-based attribution techniques applied to medical VLMs have demonstrated that the model's responses to different clinical questions activate distinct anatomical regions, confirming that the model's reasoning is grounded in relevant visual evidence. This is the primary application context of the present dissertation, where post-hoc attribution methods are applied to a generative volumetric VLM to provide spatially grounded explanations for its clinical outputs [40].

- **Genomics and Precision Medicine:** Beyond imaging, SHAP values have been applied to genomic prediction models to identify which genetic variants most strongly influence disease risk predictions, supporting the development of personalized treatment strategies.By decomposing model predictions into individual feature contributions, XAI methods help researchers understand the biological mechanisms underlying AI-driven genetic risk scores and facilitate the discovery of novel biomarkers.

- **Regulatory Compliance and Clinical Trust:** XAI methods play a critical role in meeting the regulatory requirements for AI-based medical devices under frameworks such as the EU AI Act and FDA guidelines, which increasingly mandate that AI systems used in clinical practice provide human-understandable justifications for their decisions [41]. Beyond legal compliance, XAI-generated explanations have been shown to improve clinician trust in AI systems, increase the rate of appropriate adoption of AI recommendations, and help identify cases where the model's reasoning is clinically unsound.

Despite the significant progress made in applying XAI to healthcare, the field still faces a number of unresolved technical, methodological, and practical challenges that limit the reliable deployment of explainable AI systems in clinical environments. The following challenges are particularly relevant in the context of multimodal and generative medical AI systems, which form the focus of this work.

- **Disagreement Between Methods:** Different XAI methods can produce substantially different explanations for the same model and the same input, raising fundamental questions about which explanation should be trusted. This disagreement is particularly problematic in clinical settings, where contradictory explanations from different methods could undermine clinician confidence and lead to inconsistent decision-making. The lack of a ground truth for explanations makes it difficult to determine which method is correct, motivating the need for rigorous comparative evaluation frameworks such as the one proposed in this work.

- **Computational Cost:** Some XAI methods, particularly Integrated Gradients with many interpolation steps and perturbation-based methods like SHAP, can be computationally expensive, requiring multiple forward or backward passes through a large model. This computational overhead is especially pronounced for volumetric medical models that process high-dimensional 3D data, where even a single forward pass is resource-intensive. The resulting latency may hinder real-time clinical adoption, particularly in settings where rapid diagnostic decisions are required.

- **Evaluation Difficulty:** There is no universally accepted standard for evaluating the quality of XAI explanations in medical settings. Automated faithfulness metrics such as the voxel deletion protocol measure whether highlighted regions causally drive model predictions, but they do not capture clinical utility or alignment with expert knowledge. Human evaluation through clinician studies is considered the gold standard but is expensive, time-consuming, and rarely conducted at sufficient scale to draw statistically meaningful conclusions.

- **Lack of Clinician Involvement:** Most XAI research evaluates explanations using automated metrics or proxy tasks, but few studies conduct rigorous user studies with actual

clinicians in realistic clinical workflows [42]. This disconnect between the technical development of XAI methods and their practical clinical validation means that explanations which perform well on automated benchmarks may not be useful, interpretable, or actionable for the medical professionals who are intended to use them. Bridging this gap requires interdisciplinary collaboration between AI researchers, clinicians, and human factors experts.

- **Multimodal Complexity:** The application of XAI to multimodal models introduces additional complexity, as the explanation must account for the contributions of and interactions between multiple input modalities. In vision-language models, for example, it is not sufficient to highlight important image regions alone. The explanation must also capture how the textual query modulates the model's visual attention and how cross-modal interactions shape the final output. This challenge is directly relevant to the present dissertation, which investigates how post-hoc attribution methods behave when applied to a generative volumetric vision-language model that jointly processes 3D medical scans and clinical text.

- **Scalability to High-Dimensional Data:** Medical imaging data, particularly volumetric CT and MRI scans, are inherently high-dimensional, with a single volume potentially containing hundreds of millions of voxels. Producing meaningful attribution maps over such high-dimensional inputs requires methods that can scale efficiently while preserving spatial precision. Many XAI methods were originally designed for 2D natural images and require significant adaptation to handle the volumetric nature of medical data without losing spatial resolution.

Taken together, these challenges reveal that despite the growing body of XAI research in healthcare, significant gaps remain particularly in the context of multimodal and generative architectures that operate on high-dimensional volumetric data. The following sections examine these limitations more critically and articulate the specific problem that this work addresses.

2.6 Related Works

This section reviews the most relevant existing work on the application of XAI methods to medical imaging and vision-language models. The literature is organized along three progressively more specific axes that converge toward the precise architectural and diagnostic setting addressed by this work.

2.6.1 XAI for Unimodal Medical Image Classification

The earliest and most extensively studied application of XAI in medical imaging concerns unimodal classification models, where a single imaging modality is processed by a convolutional neural network (CNN) classifier to produce a discrete diagnostic label.

Foundational Gradient-Based Saliency Early foundational work introduced saliency maps as one of the first gradient-based attribution methods, demonstrating that the gradient of a classification score with respect to the input image could highlight diagnostically relevant pixel regions [32]. Subsequent research proposed Grad-CAM, which produces coarser but more class-discriminative heatmaps by weighting convolutional feature maps by their upstream gradient importance [27]. Both methods were rapidly adopted across clinical domains, with applications to chest X-ray classification, skin lesion diagnosis, and retinal fundus analysis.

Axiomatic Scaling and Faithfulness Benchmarks Later developments introduced Integrated Gradients as a theoretically principled alternative, offering axiomatic guarantees of completeness and sensitivity that vanilla saliency mapping lacks [33]. To evaluate these visual explanations, the pixel deletion protocol was established as a quantitative framework for evaluating the causal faithfulness of attribution maps, creating a model-agnostic benchmark that remains widely used in the field [36].

Multi-Modal Cross-Evaluation More recently, several post-hoc attribution methods were systematically evaluated on multi-modal medical image analysis tasks, demonstrating their applicability across different imaging modalities and task formulations [40]. These pioneering works established the technical foundations of XAI for 2D clinical imaging, providing baseline methods for analyzing unimodal model behavior.

2.6.2 XAI for Medical Vision-Language Models

The growing adoption of multi-modal vision-language models (VLMs) in clinical settings has motivated deep research into their underlying explainability.

Attribution on Contrastive Frameworks Prior research investigated the application of post-hoc attribution methods to MedCLIP, a contrastive medical VLM, demonstrating that gradient-based explanations can significantly improve a clinician's functional understanding of the model's visual reasoning [43]. This work showed that established attribution methods can

be successfully adapted to the medical VLM setting without modifying the model's internal parameters, confirming the viability of post-hoc explainability for multimodal medical AI.

Intrinsic Attention Explanations Other works have explored internal attention-based visualization as a form of explanation for medical report generation models, utilizing the attention weights of a Transformer-based decoder to serve as a proxy for the model's visual focus during text generation [4, 30]. This approach offers a computationally efficient and easy-to-implement strategy for mapping token generation back to specific visual regions within the image space.

2.6.3 XAI for Generative and Volumetric Settings

Recent research has begun exploring the intersection of post-hoc explainability, autoregressive text generation, and high-dimensional volumetric data.

Backpropagation Through Generative Paths Methodologies mapping backpropagation through language generation layers rely on establishing gradients over the output logit space during the auto-regressive generation process. These approaches trace feature attribution by analyzing the sequential relationships between individual generated text tokens and the input pixel space.

State-of-the-Art Architectures On the architectural side, Med3DVLM represents a prominent milestone in generative volumetric medical VLMs, processing full 3D CT and MRI volumes alongside clinical text to generate free-text diagnostic responses [24]. This system demonstrates the feasibility of combining 3D convolutional backbones with multi-modal fusion mechanisms to handle high-dimensional spatial data alongside clinical text sequences.

2.7 Discussion: Limitations of Existing Work

The preceding sections have reviewed the current state of the art in multimodal deep learning, vision-language models, and explainable AI in healthcare. While significant engineering progress has been made across these domains, a critical examination of the existing literature reveals clear boundaries that motivate the present dissertation. This section synthesizes these limitations into four key structural gaps that collectively define the frontier of XAI research for multimodal medical AI systems.

2.7.1 Dominance of 2D and Unimodal Settings

The vast majority of XAI research in medical imaging remains confined to 2D and unimodal settings, where an algorithm analyzes a single static modality, such as a chest X-ray or a dermoscopic slide, processed by a standard classification backbone [27]. While these studies provide valuable insights into gradient behavior in controlled settings, their findings do not scale to the volumetric and multimodal architectures increasingly deployed in modern clinical practice.

The High-Dimensional Complexity of 3D Data Volumetric medical data, such as CT and MRI scans, introduces a depth dimension that fundamentally alters the nature of the attribution problem. A 3D attribution method must compute and project a relevance map across height, width, and depth slices simultaneously. This raises distinct challenges regarding how to select, isolate, and display the most informative cross-sections of a volumetric tensor to a clinician [24].

Computational Bottlenecks of Volumetric Encoders Furthermore, the high memory demands of processing 3D volumes impose severe practical constraints on attribution feasibility. Backpropagating multiple times through resource-heavy volumetric layers, as required by axiomatic methods like Integrated Gradients, demands immense computational overhead [33]. Existing 2D XAI benchmarks cannot handle these volumetric constraints without significant methodological adaptation, yet such frameworks are missing from current literature.

2.7.2 Focus on Classification Over Generation

The overwhelming majority of XAI methods for medical imaging have been engineered around discriminative models, specifically classifiers that map a fixed input tensor to a discrete, categorical output label [30].

The Scalar Differentiation Paradigm Popular post-hoc approaches were fundamentally designed with a scalar classification output in mind. In these settings, the attribution target is a single class score that can be cleanly differentiated with respect to the input pixels using standard backpropagation paths.

The Autoregressive Non-Differentiability Problem Generative models, by contrast, produce variable-length sequences of tokens autoregressively. Each sequential decoding step conditions on a stochastic sampling process and previously generated tokens, making the full

computation graph non-differentiable end-to-end [44]. Defining a meaningful, stable, and differentiable attribution target for autoregressive text generation is a non-trivial mathematical problem that has received very limited attention in medical XAI. Instead, existing report generation systems rely heavily on internal attention visualizations, which lack the causal faithfulness guarantees of true gradient-based attribution frameworks [36].

2.7.3 Lack of Faithfulness Evaluation in Medical VLMs

Even in niche settings where explainability has been extended to medical vision-language models, rigorous quantitative evaluation of the resulting explanations remains notably absent [43].

The Subjectivity of Visual Inspection Most existing studies evaluate explanations purely qualitatively, relying on the visual inspection of highlighted heatmaps by human raters. This approach is inherently subjective, prone to confirmation bias, and fails to prove that the highlighted pixel regions actually drove the model's inner mathematical decision [25].

The Need for Causal Benchmarking Quantitative evaluation frameworks, such as the voxel deletion protocol, offer an objective, model-agnostic measure of causal faithfulness [36]. If a post-hoc method truly identifies critical diagnostic features, progressively masking those targeted voxels must result in a steep, monotonic drop in model confidence. While this causal testing is standard in general computer vision, it has not been adapted to generative volumetric vision-language models, leaving a critical validation gap in clinical AI.

2.7.4 The Gap in Volumetric Multimodal XAI

The convergence of the three limitations described above defines the precise gap addressed by this work: there is currently a lack of systematic research applying post-hoc gradient-based attribution methods to generative volumetric vision-language models within the medical domain.

The Disconnect Between Literature and Practice Existing medical XAI work addresses either isolated 2D unimodal classifiers or 2D contrastive frameworks. However, the expansion into 3D volumetric architectures that generate free-text clinical responses remains an open frontier. This mismatch is a major barrier to real-world clinical deployment. In daily workflows, radiologists interact heavily with 3D volumes like CT and MRI scans, meaning the AI tools

most likely to offer real utility are precisely those capable of processing volumetric data and generating structured diagnostic narratives [24].

Implications for Clinical Trust If these generative multimodal systems cannot explain their text outputs in a spatially precise and causally faithful manner across all three dimensions, their clinical adoption will remain limited due to safety concerns [42]. Resolving this gap is not just an academic exercise; it is a practical necessity for the safe, transparent, and responsible deployment of generative artificial intelligence in medicine.

2.8 Motivation and Problem Statement

Having identified the key limitations of existing work in the preceding section, this section articulates the scientific and clinical motivations that drive the present dissertation and formally states the research problem that it addresses. The motivation for this work emerges from the convergence of three independent but deeply interconnected forces: the rapid advancement of generative volumetric vision-language models in medical imaging, the critical need for transparency and trustworthiness in clinical AI systems, and the absence of rigorous XAI evaluation frameworks for this class of models [4].

2.8.1 Scientific Motivation

The emergence of generative volumetric vision-language models such as Med3DVLM represents a significant leap forward in medical AI capabilities, enabling the joint processing of high-dimensional 3D medical volumes and natural language clinical queries to generate free-text diagnostic responses. However, this increase in capability comes at the cost of a corresponding increase in architectural complexity, making these models significantly more difficult to interpret and explain than their unimodal predecessors.

From a scientific standpoint, it remains unclear how post-hoc gradient-based attribution methods behave when applied to this new class of models. The cross-modal interactions between the visual encoder, the multimodal projector, and the language model decoder introduce attribution dynamics that have not been studied in existing work. Furthermore, the autoregressive generation process introduces a specific technical challenge: the stochastic token sampling at each decoding step breaks the differentiability of the computation graph, making it impossible to directly backpropagate gradients from the model's output to its input without modification. Addressing this challenge requires a principled methodological contribution that renders the generation process amenable to gradient attribution, which constitutes one of the core technical contributions of this work.

Finally, the absence of rigorous quantitative evaluation in existing medical VLM explainability work motivates the need for a systematic comparative study that goes beyond qualitative heatmap inspection. By applying a voxel deletion faithfulness protocol to four gradient-based attribution methods in a unified experimental framework, this work provides the first objective comparison of attribution method faithfulness in the generative volumetric medical VLM setting.

2.8.2 Clinical Motivation

The deployment of generative medical AI systems in real-world radiology environments is contingent on the ability of clinicians to trust and verify the model's outputs. A model that generates a diagnostic report without providing any indication of which visual evidence supports its conclusions is, from a clinical standpoint, a black box that cannot be safely integrated into diagnostic workflows. Radiologists are trained to justify their diagnoses with reference to specific anatomical findings, and they reasonably expect AI systems to meet the same standard of transparency.

The clinical motivation is further strengthened by the regulatory landscape surrounding medical AI. Frameworks such as the EU AI Act classify medical diagnostic systems as high-risk AI applications, requiring that they provide human-understandable explanations for their decisions. In this context, XAI is not merely a desirable feature but a regulatory prerequisite for clinical deployment. Beyond regulatory compliance, XAI-generated explanations can actively improve clinical decision-making by helping clinicians identify cases where the model's reasoning is clinically unsound and build calibrated trust in AI recommendations.

This clinical motivation also extends to the platform developed as part of this work. By integrating gradient-based attribution maps directly into a specialized medical imaging viewer that allows clinicians to upload CT, MRI, and X-ray scans and interact with the AI system through natural language queries, the platform transforms the theoretical contributions of this work into a practically deployable tool that bridges the gap between XAI research and clinical application.

2.8.3 Problem Statement

Building upon the scientific and clinical motivations outlined above, modern generative volumetric vision-language models represent the current state of the art in multimodal medical AI. These architectures are capable of jointly processing high-resolution, three-dimensional CT and MRI scans alongside natural language clinical queries to produce detailed, free-text diagnostic responses.

The Core Technical Bottleneck Despite their immense clinical potential, these advanced models remain fundamentally opaque black boxes. Their underlying autoregressive language generation process is inherently stochastic and non-differentiable across the sequential token output space. This mathematical constraint makes it impossible to apply standard, out-of-the-box gradient-based attribution methods without significant methodological adaptation. Furthermore, no prior research has systematically deployed post-hoc XAI frameworks on this class of volumetric generative architectures, and rigorous quantitative faithfulness benchmarks remain entirely absent from the literature.

Proposed Objective and Scope This work directly addresses this critical validation gap by adapting, implementing, and evaluating four distinct post-hoc gradient-based attribution methods on a frozen generative volumetric vision-language model baseline. To provide a robust and comprehensive evaluation, the four chosen algorithms are strategically selected to span two highly complementary interpretive families:

- **Input-Level Attribution Methods:** Comprising *Input-Level Saliency* and *Integrated Gradients*, these approaches compute gradients directly with respect to the raw input space, yielding high-resolution relevance maps at the individual voxel level.
- **Feature-Level Attribution Methods:** Comprising *Layer Grad-CAM* and *Guided Grad-CAM*, these approaches target intermediate deep convolutional activations within the 3D vision backbone, generating coarser, class-discriminative spatial heatmaps

Causal Validation Framework Rather than relying on subjective human visual inspection, the fidelity of the resulting 3D attribution volumes is rigorously measured using a quantitative, multi-positional voxel deletion protocol [36]. This framework systematically masks the highly ranked diagnostic regions to observe the corresponding decay in token generation confidence, establishing an objective, model-agnostic benchmark for explanation quality in generative clinical AI.

2.9 Conclusion

This chapter has established the theoretical and conceptual framework for explainable artificial intelligence in healthcare. The foundational concepts of XAI were defined and organized into a rigorous taxonomy, and the specific post-hoc gradient-based attribution methods relevant to this work were described in detail. The applications and challenges of XAI in healthcare were surveyed, and a critical examination of the existing literature revealed a significant gap: no prior work has systematically evaluated post-hoc attribution methods on generative volumetric

vision-language models with rigorous quantitative faithfulness assessment. Med3DVLM was identified as the ideal target model to address this gap, given its state-of-the-art 3D volumetric architecture and its direct relevance to real-world clinical radiology. The following chapter presents the methodology developed to apply and evaluate post-hoc attribution methods within this challenging setting.

Chapter 3

Conception of the System

3.1 Introduction

This chapter presents the methodological framework underlying the proposed XAI solution for generative volumetric medical vision-language models, detailing the research conception, objectives, and architectural choices that define the approach.

The chapter begins by outlining the core work objectives, followed by the global system architecture divided into three strata: the data layer, the AI model layer, and the end-user layer. It then shifts to the functional architecture, presenting the Med3DVLM model and detailing how its visual encoder, multimodal projector, and language model decoder interact. Following this, the core methodological contribution is introduced: the two-pass XAI module, explaining how teacher forcing enables gradient-based attribution map generation in a 3D volumetric setting. Finally, the quantitative evaluation protocol is outlined, covering the voxel deletion process and sample selection criteria, before concluding with the design analysis of the end-user layer.

3.2 Work Objectives

The primary objective of this work is to investigate the behavior of post-hoc gradient-based attribution methods when applied to a frozen generative volumetric vision-language model, and to evaluate their ability to produce spatially precise and causally faithful explanations for clinical outputs. To achieve this overarching goal, the work is organized around four concrete objectives:

1. **Differentiable Attribution Framework:** To design and implement a differentiable attribution framework that renders the stochastic autoregressive generation process of Med3DVLM amenable to gradient-based attribution, without modifying the model's weights or architecture. This involves introducing a teacher-forcing strategy and defining a scalar attribution target based on the cumulative log-likelihood of the model's generated

response.

2. **Systematic Attribution Method Comparison:** To systematically apply and compare four gradient-based attribution methods, namely Input-Level Saliency, Integrated Gradients, Layer Grad-CAM, and Guided Grad-CAM, within a unified experimental framework that spans multiple clinical question types, including closed-ended VQA, open-ended VQA, and radiology report generation.

3. **Quantitative Faithfulness Evaluation:** To quantitatively evaluate the faithfulness of each attribution method using a voxel deletion protocol that measures the causal impact of the highlighted regions on the model's output confidence, providing an objective and model-agnostic benchmark for comparing attribution quality.

4. **Clinical Platform Development:** To design and implement an interactive medical imaging platform that integrates the proposed attribution framework into a specialized clinical interface, allowing users to upload CT, MRI, and X-ray scans, query the AI model through natural language, and receive diagnostic insights alongside gradient-based attribution maps overlaid on the original volumetric data.

3.3 System Architecture

The global system architecture is structured into three decoupled, interacting layers that streamline the end-to-end pipeline from raw multi-modal clinical inputs to explainable visual insights. As illustrated in Figure 3.1, the system segregates its components into a Data Layer, an AI Model Layer, and an End-User Layer to balance high-throughput volumetric processing with localized, interactive client rendering.

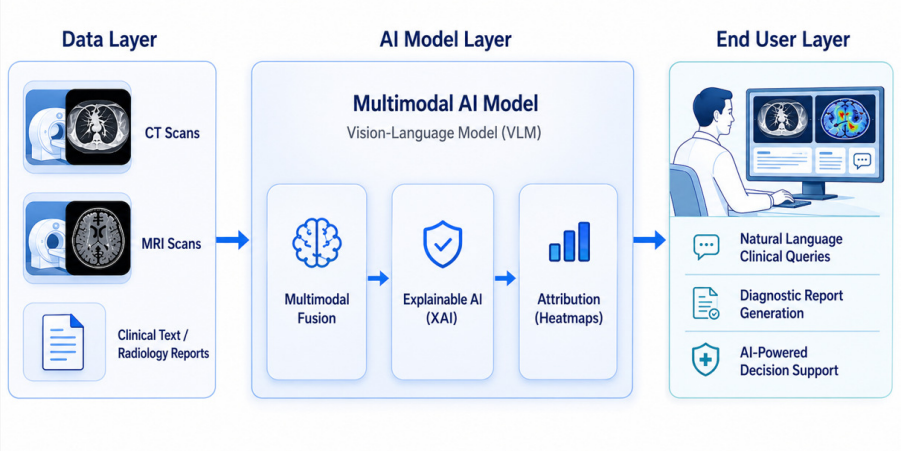


Figure 3.1: Global system architecture.

3.3.1 Data Layer

This layer manages the ingestion and synchronization of the heterogeneous, multi-modal clinical data streams that drive the system:

- **Imaging Data:** Consists of high-dimensional volumetric and spatial clinical imaging modalities, including multi-slice Computed Tomography (CT) scans and structural Magnetic Resonance Imaging (MRI) volumes.
- **Textual Data:** Encompasses unstructured clinical text sources, such as historical radiology reports, clinician findings, and dynamic free-text diagnostic queries submitted by the user.

3.3.2 AI Model Layer

The core computational engine of the architecture unifies advanced multimodal processing with post-hoc explainability functions:

- **Vision-Language Module:** Integrates visual and textual features within a shared embedding space, allowing the system to perform complex cross-modal reasoning and generate textual clinical findings conditioned on input medical scans.
- **Explainable AI (XAI) Module:** Computes feature-level and input-level attribution values directly from the model's inner representations. This module identifies the specific anatomical regions that heavily influenced the generated textual output, mapping linguistic tokens back to spatial coordinates.

3.3.3 End-User Layer

This tier serves as the primary interactive portal, delivering a dedicated, user-friendly platform designed to bridge the gap between complex model outputs and clinical users through several core capabilities:

- **Multi-Modal Visualization:** Renders contiguous medical imaging slices alongside generated clinical texts, enabling users to review visual data and textual findings in a synchronized interface.
- **Interactive Querying:** Allows users to submit open-ended diagnostic prompts to the underlying model, facilitating interactive clinical visual question answering. **Visual At-**

tribution Overlays: Projects computed explainability heatmaps directly onto the medical scans, highlighting exactly which areas the model focused on to help users evaluate the clinical reliability of the system's decisions.

3.4 Functional Architecture

The functional architecture defines the end-to-end computational pipeline, mapping raw multi-modal inputs to finalized diagnostic texts and corresponding volumetric explanation maps. As illustrated in Figure 3.2, the execution workflow passes sequentially through five core structural phases: data ingestion within the input space, feature extraction and token alignment inside the vision-language model, gradient computation via the explainable AI module, dimensional volume refinement during post-processing, and the parallel rendering of the final output modalities.

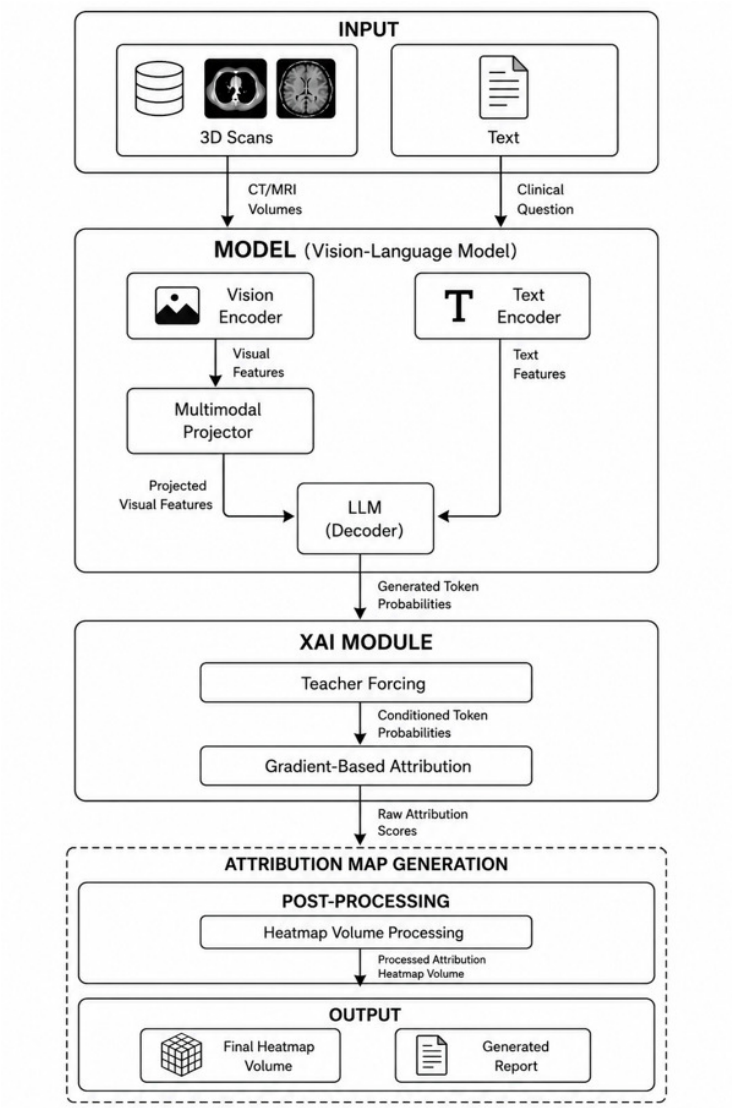


Figure 3.2: Functional computational pipeline of the explainable vision-language system.

3.4.1 Presentation Of The Model

3.4.1.1 Architecture Overview

The target model used in this work is Med3DVLM [24], a state-of-the-art vision-language model designed for three-dimensional medical image understanding. Med3DVLM is selected as the primary case study because it represents one of the most advanced generative volumetric medical VLMs currently available, combining a specialized 3D visual encoder with a large language model decoder to perform tasks such as volumetric visual question answering and automated radiology report generation.

Component Breakdown across Training Phases As illustrated in **Figure 3.3**, the structural layout and interaction of the system components adapt across two distinct operational pipelines:

- **The Contrastive Alignment Phase (Left Panel):** The architecture maps the relationship between modalities by aligning 3D volumetric visual features extracted by the **DCFormer visual encoder** with clinical text representations processed by a **text encoder** (ClinicalBERT) using a matrix similarity baseline.
- **The Generative VLM Phase (Right Panel):** The system transitions into a generative pipeline where the 3D visual encoder is connected directly to a **Transformer-based large language model decoder** via a trainable **MLP-Mixer multimodal projector**, enabling the processing of clinical questions alongside volumes to generate free-text diagnostic responses [1].

Post-Hoc Experimental Constraints All experiments in this work are conducted on the frozen Med3DVLM model setup shown on the right side of **Figure 3.3**, meaning that neither the weights nor the architecture of the model are modified during the attribution process. This adheres strictly to the fundamental requirements of the post-hoc explainability paradigm, which seeks to explain the behavior of a fixed, already-trained model without altering its native predictions [42, 40].

3.4.1.2 Visual Encoder: DCFormer

The visual encoder is based on DCFormer, a convolutional architecture specifically engineered for processing high-dimensional, three-dimensional volumetric medical data [24].

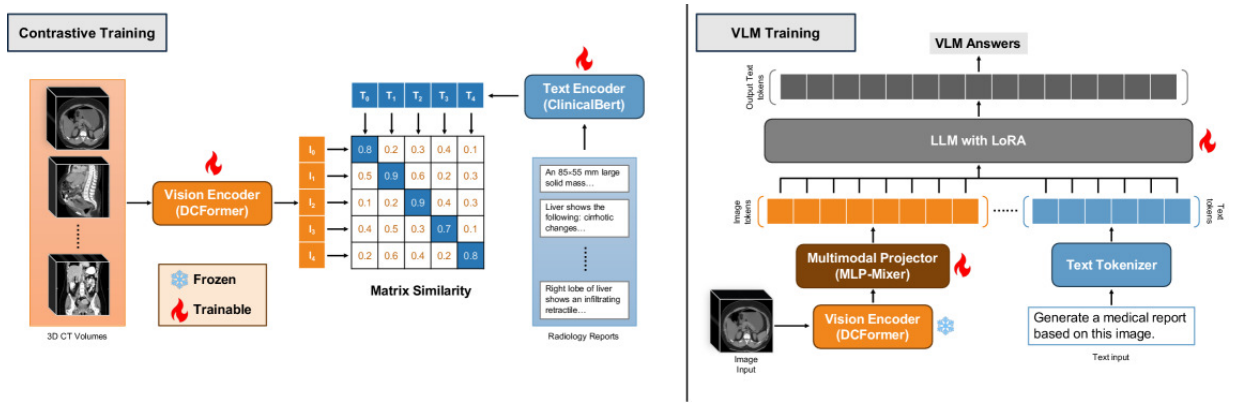


Figure 3.3: The dual-phase architecture of Med3DVLM, highlighting contrastive text alignment (left) and generative multi-modal VLM orchestration (right) [24].

Mathematical Input Formulation The encoder ingests a medical volume represented as a tensor $X \in \mathbb{R}^{C \times D \times H \times W}$, where:

- C denotes the number of input channels (typically $C = 1$ for grayscale modalities like CT or MRI).
- D denotes the structural volume depth (the total number of native slices).
- H and W denote the spatial height and width coordinates of each individual slice matrix, respectively.

To process this input, the architecture applies a series of depthwise-separable 3D convolutional blocks organized across multiple sequential stages. This layout progressively reduces the spatial dimensions (D, H, W) while increasing the latent channel depth, thereby constructing a deep hierarchical representation of the volumetric scan [45].

Target Attribution Layer Selection A key structural property of DCFormer relevant to the post-hoc attribution framework is the organization of its convolutional blocks into distinct named stages. The feature-level attribution methods investigated in this work explicitly target the specific layer:

`encoder.s4.1.dwconv.c1`

This layer corresponds to the final depthwise-convolutional block of the fourth and final encoder stage. It produces the network's deep activation maps $\mathbf{F}^{\ell} \in \mathbb{R}^{K \times D' \times H' \times W'}$, where K represents the number of feature channels, and the spatial dimensions are highly compressed ($D' \ll D$, $H' \ll H$, $W' \ll W$). These coarse activation maps serve as the mathematical target for the feature-level attribution methods described in Section 3.4.2.3.

SigLIP Vision-Language Alignment During its primary pretraining phase, DCFormer employs a SigLIP-based vision-language alignment strategy [46]. Unlike traditional contrastive objectives (such as standard CLIP) that rely on full-batch softmax normalization, SigLIP utilizes an independent, sigmoid pairwise-matching loss evaluated over image-text pairs [18].

This formulation eliminates cross-sample coupling during loss computation, significantly reducing memory overhead and improving training efficiency at scale. In practice, this design enables fast and stable optimization over large multimodal batches, forcing the encoder to yield visual tokens that are naturally aligned with clinical text spaces, which is a crucial prerequisite for effective multimodal reasoning [24].

3.4.1.3 Multimodal Projector and Decoder

The multimodal projector serves as the structural bridge connecting the 3D visual encoder to the language model decoder, projecting raw visual features into an optimal language-aligned space.

Multi-Scale MLP-Mixer Architecture As illustrated in **Figure 3.4**, the projector is implemented via a multi-scale MLP-Mixer architecture that processes data through twin parallel streams to capture multi-granularity representations. It splits features into distinct **Low Features** and **High Features** pathways, where each stream undergoes repetitive sequence mixing via N successive **Mixer Layers**.

As shown in the structural breakdown of **Figure 3.4**, each mixer block sequentially applies Layer Normalization and Multi-Layer Perceptrons (MLP) across both spatial and channel dimensions. The spatial token-mixing pathway utilizes a transposition operation ($^\top$) to map features across the token dimension, while the channel-mixing pathway processes local feature channels. Both pathways utilize an internal sequential **Fully-connected** $\rightarrow$ **GELU** $\rightarrow$ **Fully-connected** architecture. Ultimately, the multi-scale visual tokens $\mathbf{V} \in \mathbb{R}^{N_v \times d}$ are fused via a concatenation operator ($\copyright$) to project the encoder's high-dimensional spatial representations directly into the d-dimensional token embedding space of the language model, where N_v denotes the total number of visual tokens [24].

Autoregressive Transformer Decoding Pipeline The decoder is a Transformer-based large language model that generates text autoregressively, token by token. It receives a concatenated multimodal context consisting of the projected visual tokens and the tokenized clinical question, generating a coherent diagnostic answer sequence conditioned entirely on this joint multi-modal context [1, 24].

At each sequence decoding step k, the decoder maps the cumulative historical context to

produce a raw probability distribution over the entire vocabulary space $\mathcal{V}$. The system selects the target next token via a standard greedy decoding formulation:

$$\hat{a}_k = \arg\max_{v \in \mathcal{V}} \text{softmax}(\mathbf{L}_k)[v] \tag{3.1}$$

where $\mathbf{L}_k \in \mathbb{R}^{|\mathcal{V}|}$ denotes the raw logit vector produced by the decoder network at step k. This sampling process repeats sequentially until an explicit end-of-sequence token is generated, yielding the final compiled diagnostic report text $\hat{A} = (\hat{a}_1, \ldots, \hat{a}_{T_A})$ [1].

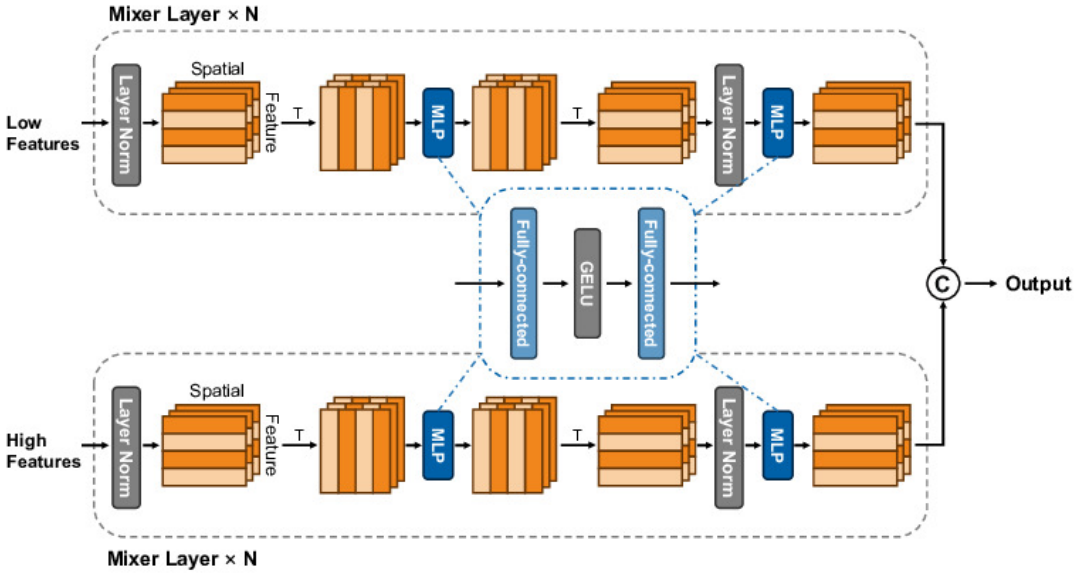


Figure 3.4: Architecture of Multi-Scale MLP-Mixer Projector

3.4.2 The XAI Module

In a standard discriminative model, gradient-based attribution is straightforward: the model produces a single scalar output, and the gradient of this scalar with respect to the input can be computed in a single backward pass. However, in a generative model such as Med3DVLM, the output is a sequence of tokens produced one at a time, where each token is sampled from a probability distribution conditioned on the previously generated tokens.

The Non-Differentiability Bottleneck This autoregressive generation process introduces two fundamental difficulties for gradient-based attribution:

- **Sampling Non-Differentiability:** The categorical sampling operation executed at each individual decoding step is inherently non-differentiable, which breaks the computation graph and prevents backpropagated gradients from flowing backward through the generation path to the input volume.

- **Stochastic Dependency Chains:** Each sequential decoding step conditions heavily on the history of previously sampled tokens. This creates a highly dynamic dependency chain that makes the exact generation path non-deterministic across multiple evaluation runs [36].

Methodological Adaptation: Dual-Pass Execution Flow To handle the non-differentiable and stochastic nature of autoregressive text generation outlined in the problem formulation, the proposed framework introduces a structured, dual-pass routing routine. This process converts the stochastic generation path into a deterministic, differentiable computation graph. As illustrated in **Figure 3.5**, the execution pipeline is formalized as follows:

The XAI Module takes the generated answer $\hat{A}$ from Med3DVLM and runs two internal passes. Pass 1 replays the autoregressive inference to obtain the full answer sequence $\hat{A}$. Pass 2 fixes that answer and re-runs a deterministic forward pass via teacher forcing, converting the stochastic generation path into a differentiable computation graph from which the scalar target S and its gradient $\partial S/\partial X$ are derived [44]. That gradient is then independently applied to each of the four attribution methods sequentially Saliency, Integrated Gradients, Layer Grad-CAM, and Guided Grad-CAM, each producing its own 3D relevance map $\mathbf{M}$.

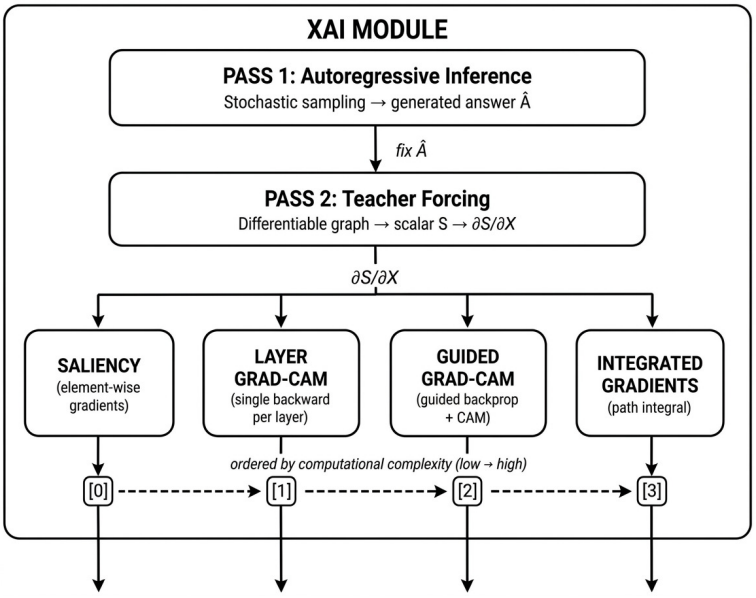


Figure 3.5: Execution Flow of the Dual-Pass XAI Module.

3.4.2.1 Pass 1: Autoregressive Generation

Let the input 3D medical volume be represented as the tensor formulation:

$$X \in \mathbb{R}^{C \times D \times H \times W} \tag{3.2}$$

where C, D, H, and W denote the channels, depth, height, and width respectively. Let $\mathcal{V}$ represent the tokenizer vocabulary, Q be the natural-language clinical query, and $\hat{A} = (\hat{a}_1, \ldots, \hat{a}_{T_A})$ (with each individual token $\hat{a}_k \in \mathcal{V}$) be the complete answer sequence generated by the model.

Multimodal Context Execution As illustrated on the left side of **Figure 3.6**, the visual and textual modalities are processed in parallel streams before converging at the language decoder. The high-dimensional volume X is mapped by the 3D visual encoder and translated into the text embedding space via the multi-scale projector, yielding a compressed visual token prefix sequence $V \in \mathbb{R}^{N_v \times d}$. Concurrently, the text tokenizer $\tau(\cdot)$ maps the clinical query Q to discrete vocabulary indices $I_Q = \tau(Q)$. These two representation streams are joined to form a single multimodal context vector:

$$I = [V,\ I_Q] \tag{3.3}$$

Within this unified context, V occupies the leading positions as the visual prefix to establish the diagnostic state, while I_Q follows immediately after to dictate the focus of the question.

Autoregressive Inference Loop This combined context vector feeds into the language decoder, which predicts the subsequent answer tokens sequentially. At each decoding step k, the network computes a logit array $L_k \in \mathbb{R}^{|\mathcal{V}|}$ and selects the next target token via standard greedy inference:

$$\hat{a}_k = \arg\max_{v \in \mathcal{V}} \ \mathrm{softmax}(L_k)[v] \tag{3.4}$$

This autoregressive cycle loops continuously until an explicit end-of-sequence token is generated, compiling the complete diagnostic text sequence $\hat{A}$. Because this execution pathway relies on discrete sampling choices, it forms a non-differentiable graph where backpropagation cannot pass through.

3.4.2.2 Pass 2: Teacher Forcing and Scalar Target

To bypass the sampling non-differentiability established during inference, the pipeline initiates a second evaluation block. As shown in the transition across **Figure 3.6**, the generated text sequence $\hat{A}$ is instantly frozen and routed into a teacher-forced routine [44].

Deterministic Graph Transformation During this teacher-forced pass, the model re-evaluates the identical input volume X and user question Q. However, instead of allowing the network to sample tokens dynamically, each decoding step k is fed the fixed history prefix $\hat{a}_{<k} = (\hat{a}_1, \ldots, \hat{a}_{k-1})$ directly from the frozen answer string. This intervention transforms the stochastic inference loop into a fully deterministic forward pass. Because the discrete token

sampling steps are bypassed entirely, the operational pipeline becomes an end-to-end differentiable computation graph.

Gradient Extraction and Routing Path This continuous graph allows backpropagation signals to flow uninterrupted from the decoder's output layer back to the raw input features. The system computes a scalar target score $S(X, Q, \hat{A})$ by tracking the cumulative logit values of the generated tokens. As traced by the backward path ($\frac{\partial S}{\partial X}$) on the right side of **Figure 3.6**, this stable mathematical target is differentiated to yield a base gradient tensor. This clean gradient signal is subsequently directed to the four individual post-hoc attribution modules to compute the final spatial relevance maps $\mathbf{M}$.

Using teacher forcing over alternative estimation tricks (such as straight-through approximations) ensures strict path-faithfulness: the resulting gradient traces the model's actual clinical output path rather than a proxy mathematical approximation [33].

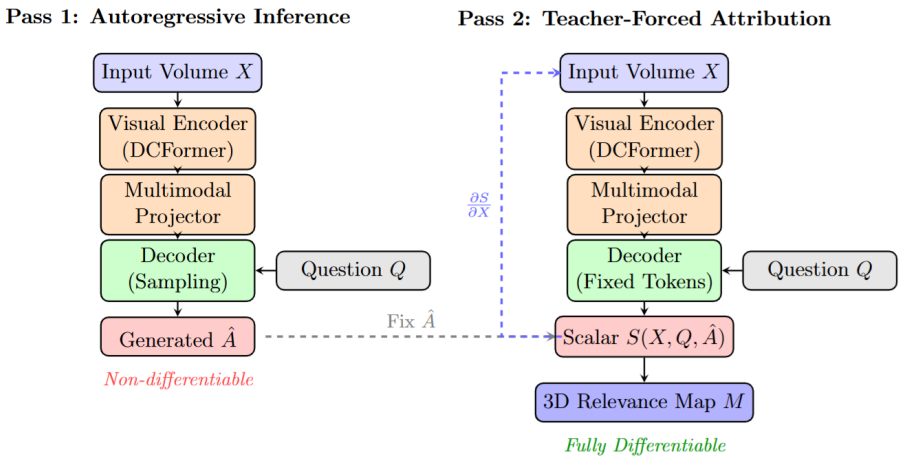


Figure 3.6: Two-Pass Attribution Framework Architecture

Scalar Attribution Target The attribution target is defined as the cumulative log-likelihood of the model's own previously generated response. This target is chosen because autoregressive decoders are originally trained to maximize this same sequence log-likelihood (equivalently, to minimize negative log-likelihood), so the attribution objective directly matches the model's primary optimization goal [44]:

$$S(X, Q, \hat{A}) = \sum_{k=1}^{T_A} \log p(\hat{a}_k \mid X, Q, \hat{a}_{<k}), \tag{3.5}$$

where T_A is the total number of generated answer tokens, $k \in \{1, \ldots, T_A\}$ indexes each decoding step, $\hat{a}_k \in \mathcal{V}$ is the k-th token of the previously generated answer $\hat{A}$, $\hat{a}_{<k} = (\hat{a}_1, \ldots, \hat{a}_{k-1})$ denotes the teacher-forced prefix supplied to the decoder at step k, X is the input 3D volume (Eq. (3.2)), and Q is the clinical question.

The conditional probability at step k is obtained from the decoder logits L_k as:

$$p(\hat{a}_k \mid X, Q, \hat{a}_{<k}) = \text{softmax}(L_k)[\hat{a}_k]. \tag{3.6}$$

This formulation has several important properties. Summing the log-probabilities over the complete generated sequence makes S reflective of the model's joint confidence in its entire response rather than in any single token. The resulting scalar provides a single, well-defined target for gradient computation that captures how the model's overall answer depends on the input volume. Because teacher forcing eliminates all stochasticity from the forward pass, the gradients $\partial S/\partial X$ are fully reproducible across runs [33, 36].

All attribution methods described below treat the pipeline $\mathcal{M} : (X, I) \mapsto S$ as their differentiable black box and produce a 3D relevance map $M \in \mathbb{R}^{D\times H\times W}$ that can be inspected as axial, coronal, or sagittal slices [40, 24].

3.4.2.3 Attribution Methods

The four attribution methods introduced in Chapter 2 are adapted in this thesis to the specific constraints of the generative volumetric setting of Med3DVLM.

The key adaptation common to all four methods is the replacement of the standard scalar classification target with the cumulative log-likelihood objective S defined in Section 3.4.2.2, which captures the model's joint confidence in its entire generated response rather than a single class score.

All methods treat the teacher-forced pipeline $\mathcal{M} : (X, I) \mapsto S$ as their differentiable black box and produce a 3D relevance map $M \in \mathbb{R}^{D\times H\times W}$.

1. **Input-Level Saliency**
   Saliency is applied without structural modification.

   The absolute gradient of S with respect to each voxel of the input volume X is computed in a single backward pass through the teacher-forced computation graph:

   $$R_{\text{Sal}}(x_{d,h,w}) = \left| \frac{\partial S(X, Q, \hat{A})}{\partial x_{d,h,w}} \right|. \tag{3.7}$$

   The adaptation to the volumetric setting is straightforward since the method operates directly on the raw voxels of X and requires no assumption about the spatial dimensionality of the input.

2. **Integrated Gradients**

   The key adaptation of Integrated Gradients to this setting concerns the choice of baseline.

   The zero volume $X_0 = \mathbf{0}$ is used as the reference point, representing the complete absence of signal. This is a natural and semantically meaningful choice for CT inputs whose intensities are non-negative after normalization, and ensures that the completeness axiom is satisfied with respect to the scalar objective S:

$$mathrmIG(X) = (X - X_0) \odot \int_0^1 \frac{\partial S(X_0 + \alpha(X - X_0),\, Q,\, \hat{A})}{\partial X}\, d\alpha. \tag{3.8}$$

   The integral is approximated using the Riemann midpoint rule with 50 uniformly spaced interpolation steps, each requiring a separate teacher-forced forward and backward pass.

   Gradient accumulation is used to avoid storing all 50 computation graphs simultaneously, keeping memory usage within the hardware constraints.

3. **Layer Grad-CAM**

   The main adaptation of Layer Grad-CAM concerns the selection of the target layer.

   Given the 3D convolutional architecture of DCFormer, the target layer is set to the last depthwise-convolutional block of the vision tower, specifically $\ell =$ `encoder.s4.1.dwconv.c1`, with activation maps $F^\ell \in \mathbb{R}^{K \times D' \times H' \times W'}$.

   This layer is selected because it represents the deepest convolutional representation before the visual features are projected into the language model embedding space.

   The standard Grad-CAM formulation is extended to three spatial dimensions by performing global average pooling over depth, height, and width simultaneously:

$$w_k = \frac{1}{D'H'W'} \sum_{d,h,w} \frac{\partial S(X, Q, \hat{A})}{\partial F^\ell_{k,d,h,w}}. \tag{3.9}$$

   The coarse relevance map is then formed by a weighted sum across channels, followed by a ReLU to retain only positively contributing regions:

$$R_{\mathrm{LGCAM}}(X) = \mathrm{ReLU}\left(\sum_{k=1}^{K} w_k\, F^\ell_k \right). \tag{3.10}$$

   The resulting coarse 3D heatmap is upsampled via trilinear interpolation to the input resolution $D \times H \times W$, extending the original 2D bilinear upsampling to the volumetric setting.

4. **Guided Grad-CAM**

   Guided Grad-CAM is applied without modification to its fusion strategy.

The Grad-CAM component uses the same target layer ℓ as Layer Grad-CAM, and the Guided Backpropagation component backpropagates $\partial S/\partial X$ through the full teacher-forced computation graph with the modified ReLU rule.

The two signals are fused element-wise and normalized:

$$R_{\mathrm{GGCAM}}(X) = \mathrm{Norm}\big(C_\ell(X) \odot \tilde{G}_X\big). \tag{3.11}$$

The primary adaptation here is the extension of the trilinear upsampling of $C_\ell(X)$ to three spatial dimensions, ensuring that the coarse and fine-grained signals are spatially aligned before fusion.

3.4.3 Attribution Map Generation

Each attribution method independently processes the gradient signal to produce a raw, three-dimensional relevance map, where each entry quantifies the contribution of the corresponding voxel to the model's confidence in its generated tokens. For feature-level methods, the map is initially computed at the coarser resolution of the target encoder layer and upsampled to the full input resolution.

3.4.3.1 Post-Processing

As illustrated in the functional pipeline, the raw attribution scores must pass through a dedicated Heatmap Volume Processing phase before presentation. This post-processing layer handles the spatial transformation of the three-dimensional matrix through three sequential operations:

1. **Slice Selection:** The axial slice with the highest summed attribution across the 3D grid is extracted as the most informative cross-section, referred to as the top-1 slice. For feature-level methods like Layer Grad-CAM, whose diffuse activations do not yield a sharp peak slice, a consensus slice is defined using the mean top-1 index across the input-space methods to approximate the anatomically most relevant depth.
2. **Volume Normalization:** The extracted spatial slice is normalized to $[0, 1]$ via min-max scaling to ensure a consistent range for visual comparison across diverse attribution backbones.
3. **Color Rendering:** The normalized scalar values are mapped to a color spectrum (ranging from transparent blue for low relevance to intense red for high relevance) to convert the raw numerical matrices into a human-interpretable visual format.

3.4.3.2 Multi-Modal Output Generation

The pipeline terminates at the Output layer, which concurrently yields two synchronized modalities to provide a complete diagnostic decision loop:

- **Generated Report:** The finalized natural language clinical text sequence decoded by the language model in response to the clinical question.
- **Final Heatmap Volume:** The processed attribution heatmap overlaid onto the original grayscale 3D scan volume. Bright, localized clusters indicate highly localized visual explanations, whereas diffuse heatmaps reveal broader visual coverage. This dual output format explicitly links linguistic findings back to their spatial anatomical coordinates.

3.5 Evaluation Protocol

Evaluating the quality of attribution maps is a non-trivial task, as there is no ground truth for what a correct explanation looks like. Visual inspection of heatmaps is inherently subjective and insufficient for drawing rigorous scientific conclusions about the relative faithfulness of different attribution methods [36]. To address this limitation, this dissertation adopts a quantitative faithfulness evaluation protocol based on the voxel deletion paradigm, which provides a model-agnostic and objective measure of whether the regions highlighted by an attribution method are causally responsible for the model's output.

3.5.1 Voxel Deletion Protocol

The voxel deletion protocol evaluates attribution faithfulness by measuring the causal impact of progressively removing the most important voxels identified by each method on the model's output confidence. The underlying intuition is straightforward: if an attribution method correctly identifies the voxels that are most responsible for the model's decision, then progressively masking those voxels should cause a steep and monotonic drop in the scalar attribution target S. Conversely, if the highlighted voxels are not causally important, masking them will have little effect on the model's confidence, resulting in a flat or slow-decaying deletion curve [36].

The protocol is executed as follows:

1. **Rank voxels by importance:** For each attribution method, the voxels of the input volume X are ranked in descending order of their attribution scores, producing an ordered list from the most to the least important voxel .

2. **Progressively mask top-ranked voxels:** The top-k most important voxels are progressively masked by setting their intensity values to zero, evaluated at increments of 10%, 20%, and 30% of the total number of voxels. Setting masked voxels to zero is consistent with the zero baseline used in Integrated Gradients and represents the absence of signal in the volume.

3. **Re-evaluate the model:** At each deletion step, the degraded input volume is passed through the frozen Med3DVLM model using teacher forcing, and the scalar attribution target S defined in Equation (3.5) is re-computed. This ensures that the evaluation is performed under the same differentiable framework used during attribution.

4. **Analyze the deletion curve:** The resulting sequence of S values is plotted as a function of the percentage of voxels masked, producing a deletion curve for each attribution method. A steep, monotonic decrease in S confirms that the deleted voxels were causally important for the model's decision and that the attribution method is faithful. A slow or non-monotonic decay indicates that the highlighted regions are not causally responsible for the model's output, suggesting that the method lacks spatial specificity.

The deletion curves for all four attribution methods are computed on the same input sample and displayed together, enabling a direct visual comparison of their relative faithfulness. The method that achieves the steepest confidence drop at the lowest deletion percentage is considered the most faithful, as it identifies the smallest set of voxels that are collectively most responsible for the model's output.

3.5.2 Sample Selection Criteria

To ensure a meaningful and unambiguous faithfulness assessment, the evaluation is restricted to input samples on which Med3DVLM produces correct, high-confidence answers. This criterion is applied consistently across all four attribution methods to ensure a fair comparison. Samples where the model produces incorrect or low-confidence responses are excluded, as the attribution maps for such samples may reflect the model's uncertainty rather than its genuine reasoning process, making faithfulness evaluation unreliable.

The evaluation is conducted across three clinical question types, namely closed-ended VQA, open-ended VQA, and radiology report generation, to assess whether the faithfulness hierarchy between attribution methods is consistent across different task formulations or varies with the nature of the clinical query.

3.6 End-User Layer Analysis and Design

This section focuses on the analysis and design of the End-User layer. It determines the functional and non-functional needs expected from the platform, and the interactions between it and users. Then, it covers both static and dynamic aspects. Where the static aspect defines the structural requirements and user roles, which represent the functional scope of the RadEx platform. While, the dynamic aspect represents the behavior over time, including component interactions, data flow, and the sequence of operations.

3.6.1 Functional Requirements (Use Case Diagrams)

The functional scope of the platform is defined through use case diagrams for both the clinical and educational modules. Figure 3.7 illustrates the core clinical workflows, detailing the clinician's interactions with the system, including volumetric scan uploads, XAI inference generation, and the retrieval of diagnostic reports.

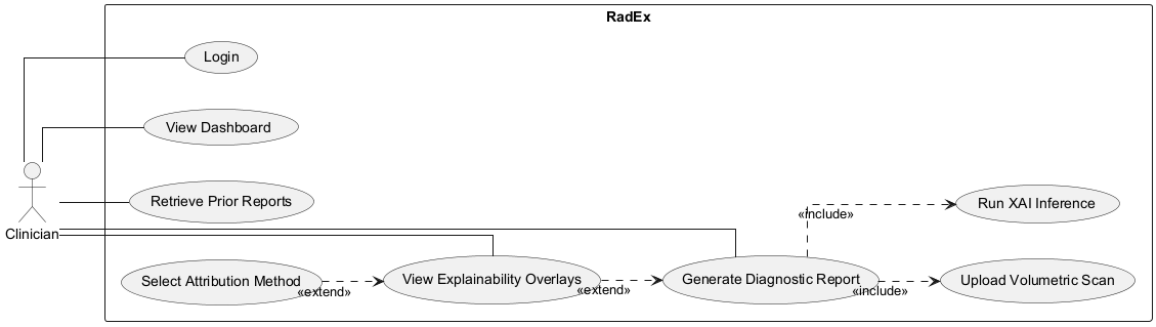


Figure 3.7: RadEx Clinical Use Case Diagram.

Similarly, Figure 3.8 presents the use cases for the educational module, capturing how radiology residents engage with practice cases, submit findings, and receive automated grading feedback.

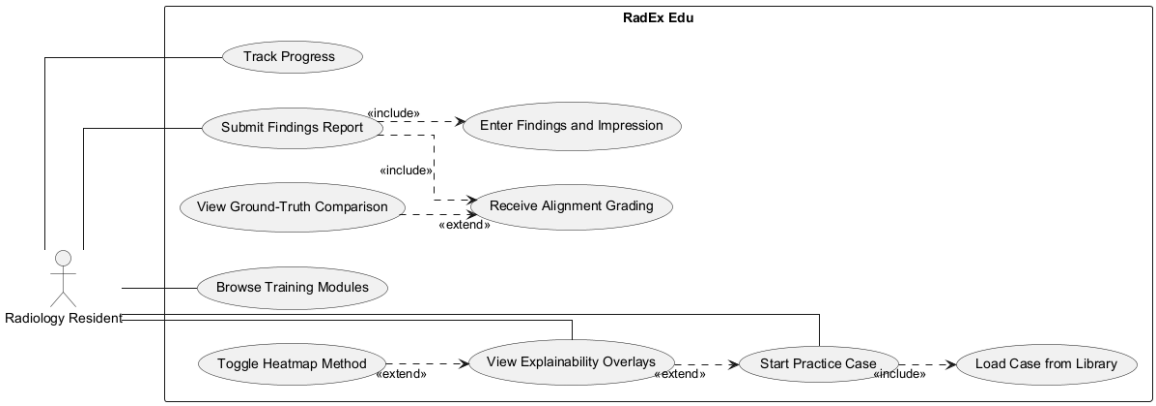


Figure 3.8: RadEx Edu Use Case Diagram.

3.6.2 Static Structure (Class Diagrams)

The structural design of the platform encompasses presentation components, business logic controllers, and core domain entities. Figure 3.9 presents the class structure for the clinical application, highlighting the orchestration of clinical studies, prompts, and the generation of diagnostic reports with their associated attribution overlays.

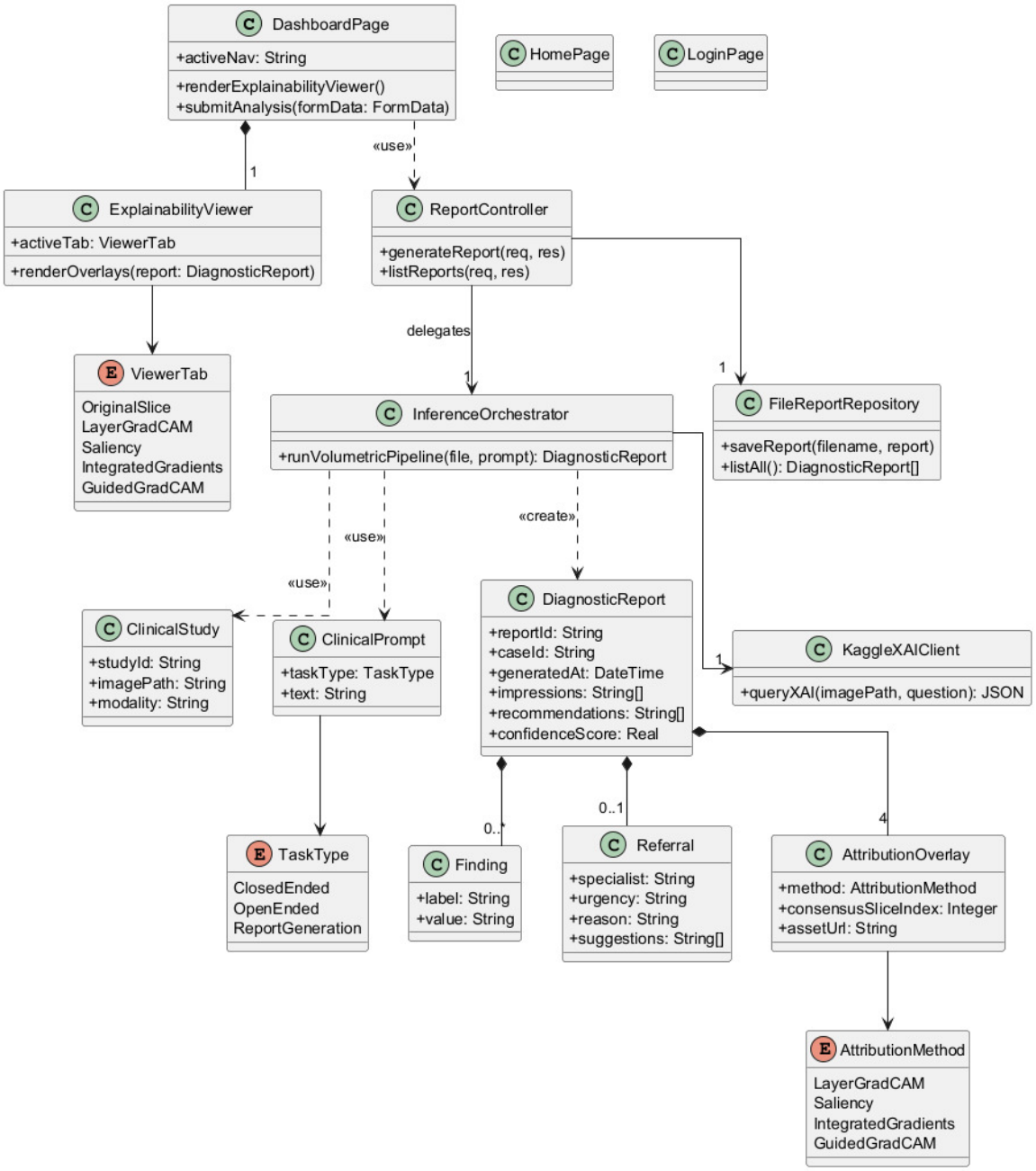


Figure 3.9: RadEx Clinical Class Diagram.

Figure 3.10 models the educational domain, delineating the relationships between training modules, resident submissions, grading engines, and progress tracking entities.

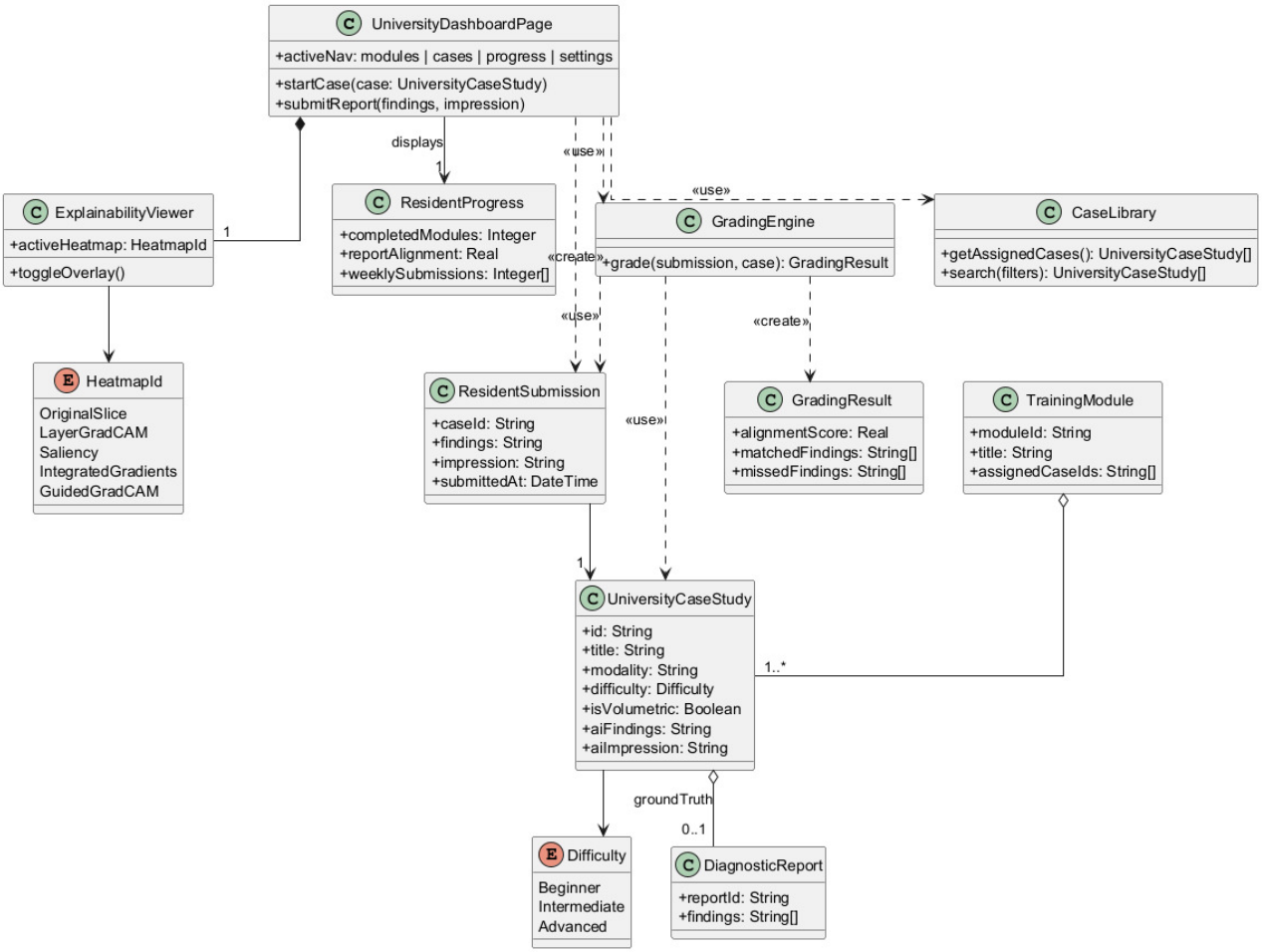


Figure 3.10: RadEx Edu Class Diagram.

3.6.3 Dynamic Behavior (Sequence Diagrams)

The dynamic sequence of operations defines how the system's components interact over time to fulfill key use cases. Figure 3.11 details the execution flow when a clinician initiates report generation, tracing the process from the client interface through the API to the remote inference engine, and culminating in the delivery of the diagnostic report.

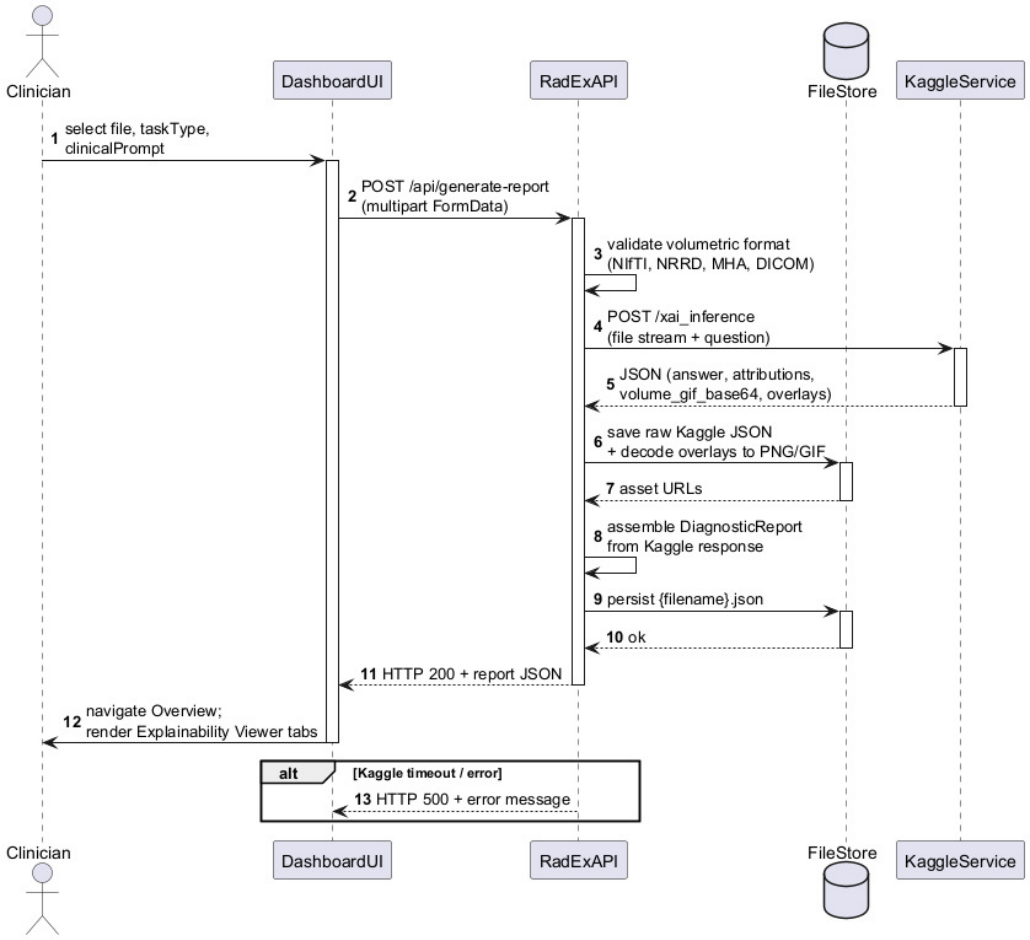


Figure 3.11: RadEx Report Generation Sequence Diagram.

Figure 3.12 depicts the sequence for retrieving previously generated reports to populate the user dashboard and report lists.

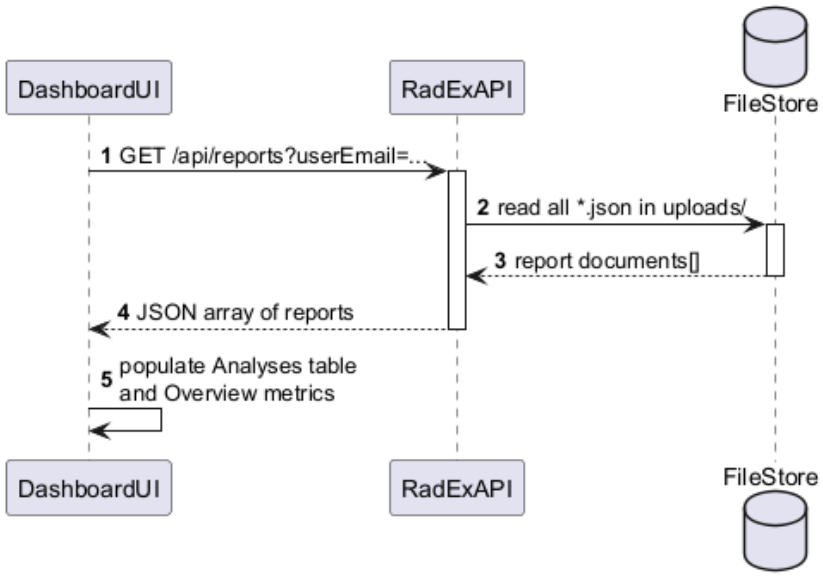


Figure 3.12: RadEx Report Retrieval Sequence Diagram.

Figure 3.13 illustrates the educational sequence, demonstrating how a resident starts a practice case, submits their impression, and receives grounded assessment feedback against the AI consensus.

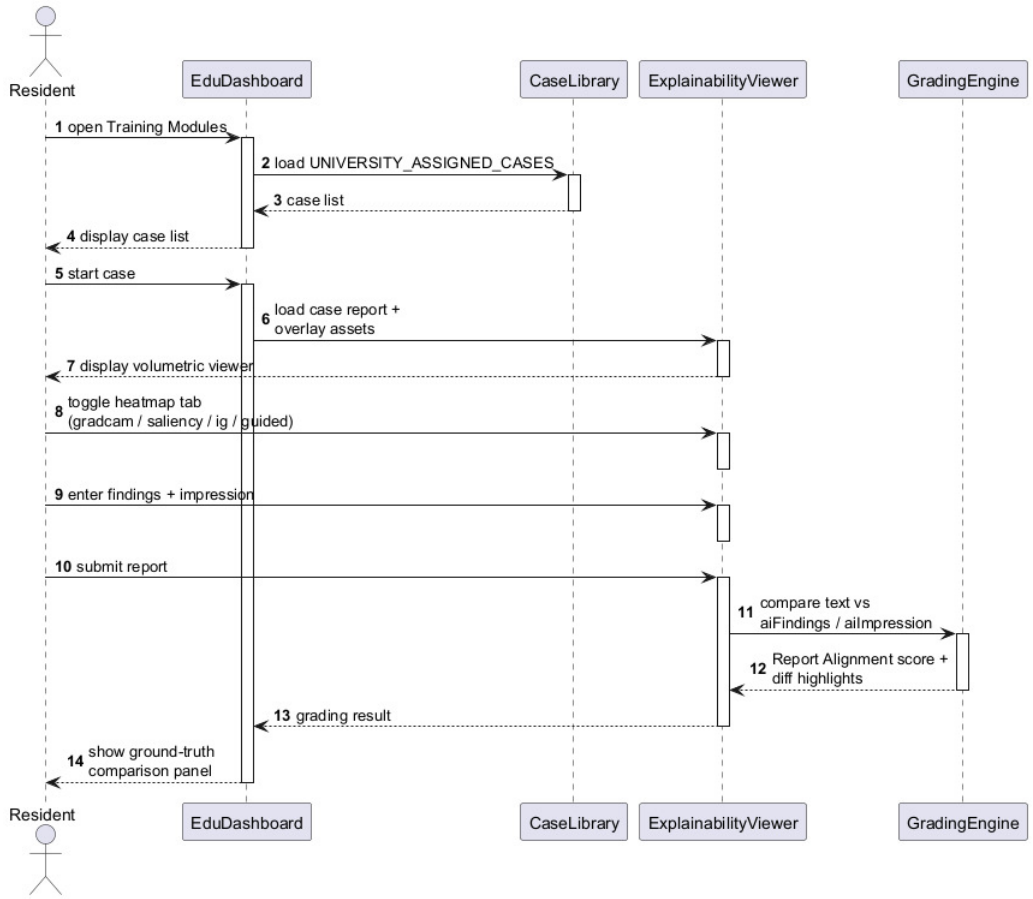


Figure 3.13: RadEx Edu Practice Sequence Diagram.

3.7 Conclusion

In summary, this chapter has laid out the complete conceptual and methodological architecture of the proposed trustworthy AI framework for 3D medical vision-language processing. By mapping out the global system design across its structural data, AI model, and end-user layers, a clear blueprint was established for translating complex multimodal clinical inputs into human-interpretable diagnostics. The breakdown of the Med3DVLM model further clarified how spatial features from the DCFormer encoder and tokenized queries from the text space are seamlessly combined via the MLP-Mixer projector before being decoded into natural language.

At the core of this methodological setup, the introduction of the dual-pass XAI module solves the fundamental challenge of non-differentiable autoregressive generation. Utilizing a teacher-forcing mechanism guarantees a stable, deterministic computation graph, which successfully unlocks the application of post-hoc gradient-based attribution to volumetric contexts.

This mathematical grounding directly underpins the deployment of the four specific attribution techniques adapted for three-dimensional medical analysis. Furthermore, the established voxel deletion protocol and rigorous sample selection criteria provide a robust quantitative foundation to verify the faithfulness of these visual maps. Finally, the deep analysis and functional design of the end-user layer bridges the gap between raw attribution values and an intuitive, clinically viable interface, setting the stage for the empirical evaluation and implementation details presented in the upcoming chapters.

Chapter 4

Implementation and Results

4.1 Introduction

Having established the methodological framework and research design in the preceding chapter, this chapter proceeds to the concrete implementation of the proposed XAI attribution pipeline and presents the experimental results obtained by applying it to Med3DVLM. The transition from methodological formulation to implementation involves translating the theoretical design into a working computational system, configuring the software and hardware environment, and executing the attribution experiments across a set of carefully selected clinical samples.

The chapter is organized into two main parts. The first part covers the implementation details, describing the dataset, software framework, hardware environment, and the technical aspects of the attribution pipeline and visualization strategy. The second part presents the experimental results, combining qualitative attribution map analysis across three clinical question types with a quantitative voxel deletion faithfulness evaluation of all four attribution methods. The chapter concludes with a comparative analysis positioning the proposed approach relative to existing XAI work in medical imaging, followed by a detailed discussion of the key findings, their clinical implications, and the limitations of the present study.

4.2 Experimental Setup

4.2.1 Dataset

Attribution is performed using volumetric CT scans from the M3D-Cap and M3D-VQA datasets. Due to copyright restrictions imposed by the dataset providers, only few volumetric images were accessible for experimentation. Each of the few volumes is queried with three distinct clinical question types: closed-ended VQA, open-ended VQA, and report generation.

The few selected volumes span CT scans covering diverse anatomical structures and patholo-

gies, ensuring that the attribution evaluation is not biased toward a single anatomical region [3].

4.2.2 Software Framework

The experimental framework is implemented in Python using the PyTorch deep learning framework. The attribution methods are implemented using the Captum library, a widely-used open-source interpretability library built on top of PyTorch. Captum provides modular, well-tested implementations of numerous gradient-based and perturbation-based attribution methods, including Saliency, Integrated Gradients, Layer Grad-CAM, and Guided Grad-CAM[47, 48].

4.2.3 Hardware Environment

All experiments are conducted within a Kaggle computing environment utilizing an NVIDIA P100 GPU with 16 GB of dedicated video memory. For Integrated Gradients, each of the 50 interpolation steps requires a separate forward and backward pass, and gradient accumulation is used to avoid storing all 50 computation graphs simultaneously.

4.2.4 Method Configuration

All four attribution methods are evaluated on the frozen Med3DVLM backbone using the scalar objective S defined in Equation 3.5 as the universal attribution target.

Evaluation Constraint To ensure a mathematically sound and meaningful causal faithfulness assessment, the evaluation protocol is strictly restricted to data samples where Med3DVLM natively produces correct, high-confidence diagnostic answers. This criterion guarantees that the attribution methods are explaining true model comprehension rather than random hallucinatory noise, and is applied consistently across all experimental runs.

Algorithm-Specific Hyperparameters The specific execution parameters for the four evaluated post-hoc XAI methods are configured as follows:

- **Input-Level Saliency:** The base spatial gradient $\frac{\partial S}{\partial X}$ is extracted via a single backward pass flowing completely through the full encoder-projector-decoder computation graph.
- **Integrated Gradients:** The path integral is numerically approximated using 50 Riemann-midpoint steps. These steps are evaluated along a linear interpolation trajectory from a zero-volume black baseline ($X_0 = \mathbf{0}$) to the original raw input volume X.

- **Feature-Level Methods (Grad-CAM variants):** For both Layer Grad-CAM and Guided Grad-CAM, the target extraction layer ℓ is explicitly bound to the final depthwise-convolutional block of the DCFormer vision tower.

Resolution Post-Processing Because the feature-level methods compute relevance on deeply compressed internal activation maps, all generated attribution volumes across all four methods are systematically upsampled via trilinear interpolation. This restores the heatmaps to the native spatial resolution (D, H, W) of the original medical input volume prior to both clinical visual display and quantitative voxel-deletion evaluation.

4.2.5 Visualization Strategy

For the three input-space methods (Saliency, Integrated Gradients, and Guided Grad-CAM), the axial slice with the highest summed attribution is selected for display. Each method independently identifies the depth index n at which the total attribution mass is maximized, and the corresponding slice is rendered with bright-red voxel clusters overlaid on the dark grayscale CT slice.

For Layer Grad-CAM, a different visualization strategy is adopted. Layer Grad-CAM's broad, diffuse activation volume does not yield a meaningful peak slice of its own: its maximum-attribution depth is dominated by anatomically prominent but informationally non-selective regions. To address this issue, the displayed slice for Layer Grad-CAM is set to the rounded mean of the top-1 slice indices from the three input-space methods, referred to as the consensus slice. The consensus slice approximates the anatomically most relevant depth as identified by the more spatially focused methods, enabling a direct visual comparison on the same anatomical cross-section. Layer Grad-CAM maps are rendered as smooth heatmaps blended over the CT slice, with white-yellow indicating higher attribution.

Table 4.1: Summary of the implementation configuration used in this dissertation.

Component	**Configuration**
Deep Learning Framework	PyTorch
Attribution Library	Captum
Computing Environment	Kaggle
GPU	NVIDIA P100 (16 GB)
IG Interpolation Steps	50
Dataset	M3D-Cap / M3D-VQA
Question Types	Closed-ended, Open-ended, Report-generation
Deletion Increments	10%, 20%, 30%, 40% ,...., 100%

4.3 Qualitative Results

This section presents a detailed qualitative analysis of the attribution maps generated by each of the four methods across the three distinct clinical question types.

4.3.1 Input-Space Methods Across Question Types

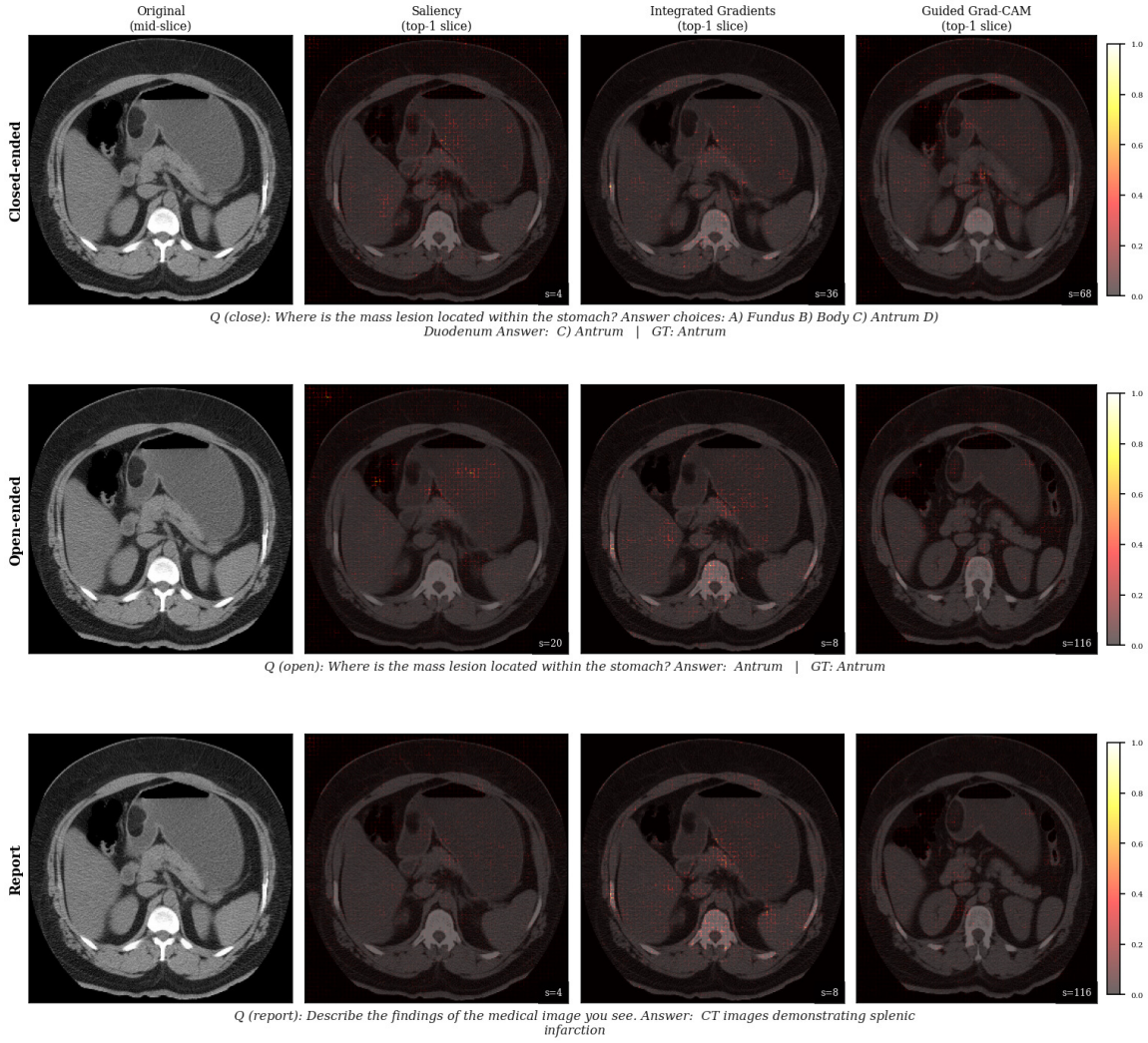


Figure 4.1: Attribution maps across VQA and report generation formats

Figure 4.1 explicitly compares the visual output for targeted VQA against open-ended report generation on the same axial slice. Visually, the input-space methods (Saliency, IG, Guided Grad-CAM) render as highly concentrated, bright red pixel clusters that physically morph and shift their spatial footprint as the question scope expands. Conversely, Layer Grad-CAM presents as a static, widespread yellow-orange thermal blob that spans the vast majority of the slice, visually oblivious to the shift in the textual clinical prompt.

4.3.1.1 Pelvic CT Case Analysis

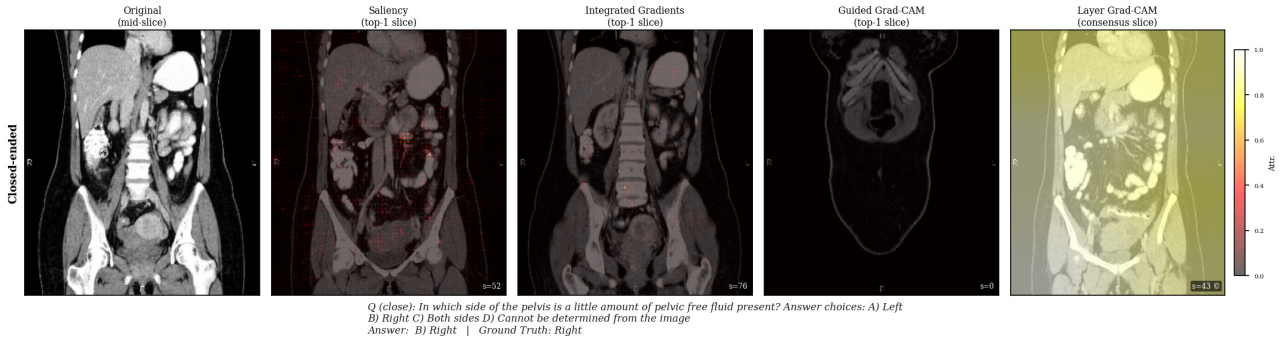


Figure 4.2: Closed-ended attribution maps for pelvic CT case

Figure 4.2 shows a pelvic CT evaluating a closed-ended targeted question about the presence of pelvic free fluid. The correct answer (B: Right) aligns with input-level methods. Saliency (slice $s = 52$) captures scattered noise, whereas Integrated Gradients ($s = 76$) precisely concentrates activation around the pelvic cavity on the right side. Guided Grad-CAM highlights sharp anatomical edges, though less centered ($s = 0$). Layer Grad-CAM ($s = 43$) broadly envelops the cross-section with an imprecise thermal band, unable to isolate the finding.

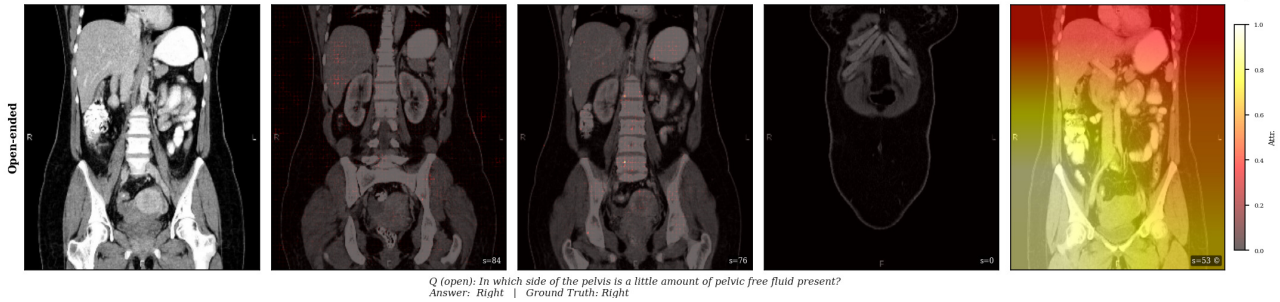


Figure 4.3: Open-ended attribution maps for pelvic CT case

Figure 4.3 illustrates the responses to an open-ended unconstrained question. Saliency shifts to deeper slices ($s = 84$) generating fine red activations near the suspected cystic area. Integrated Gradients remains highly selective at $s = 76$, reinforcing its spatial stability. Guided Grad-CAM preserves edges but remains displaced ($s = 0$). Layer Grad-CAM updates its broad attention ($s = 53$), yet still lacks diagnostic-level spatial specificity compared to the input-level techniques.

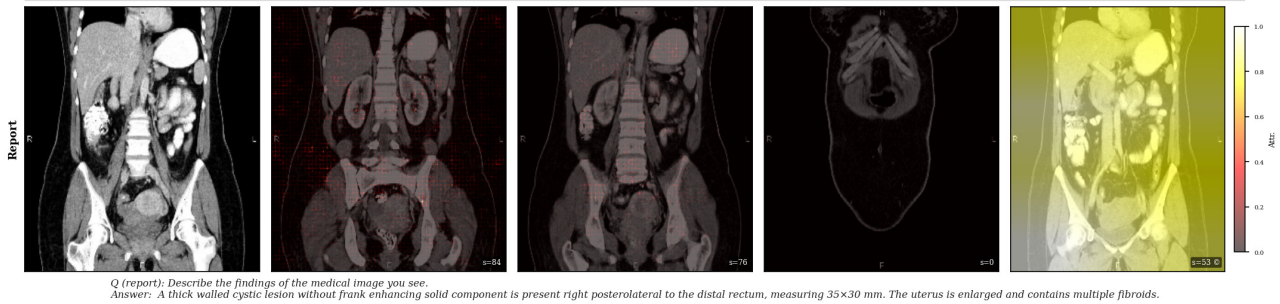


Figure 4.4: Report generation attribution maps for pelvic CT case

Figure 4.4 evaluates the attribution footprint during the generation of the descriptive clinical report. The textual output accurately records a thick-walled cystic lesion on the right. Both Saliency ($s = 84$) and Integrated Gradients ($s = 76$) respond by intensifying their focus around the right posterolateral pelvic compartment and adjacent soft tissues. Layer Grad-CAM ($s = 53$) continues to exhibit static, large-scale saturation, demonstrating a failure to adapt to localized diagnostic findings.

4.3.1.2 Abdominal CT Case Analysis

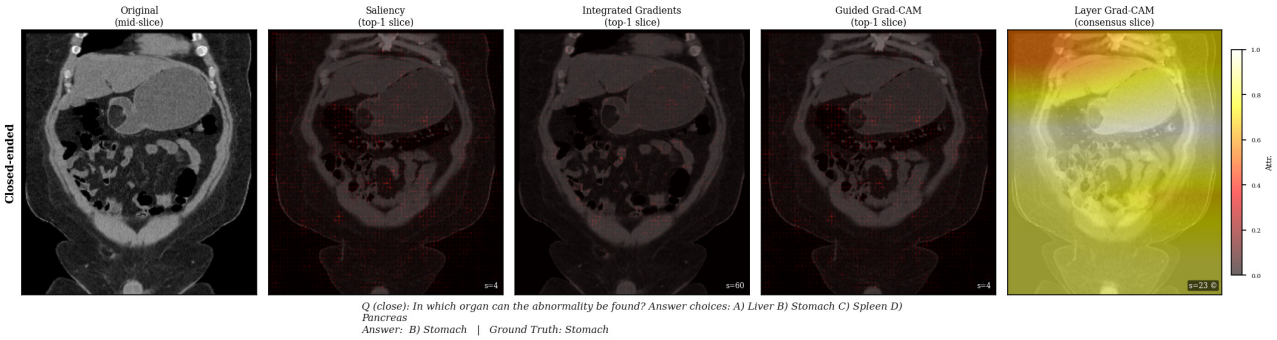


Figure 4.5: Closed-ended attribution maps for abdominal CT case

Figure 4.5 examines a closed-ended query asking which organ contains a mass. The model accurately outputs "Stomach." For this constrained reasoning task, Saliency ($s = 4$) identifies the stomach contour and boundary but suffers from scattered non-target speckle. Integrated Gradients ($s = 60$) offers the cleanest localization, selectively emphasizing the gastric wall and the hypodense lesion itself. Layer Grad-CAM forms sweeping bands across the mid-abdomen ($s = 23$) indicating systemic organ-level recognition but lacking specific local granularity.

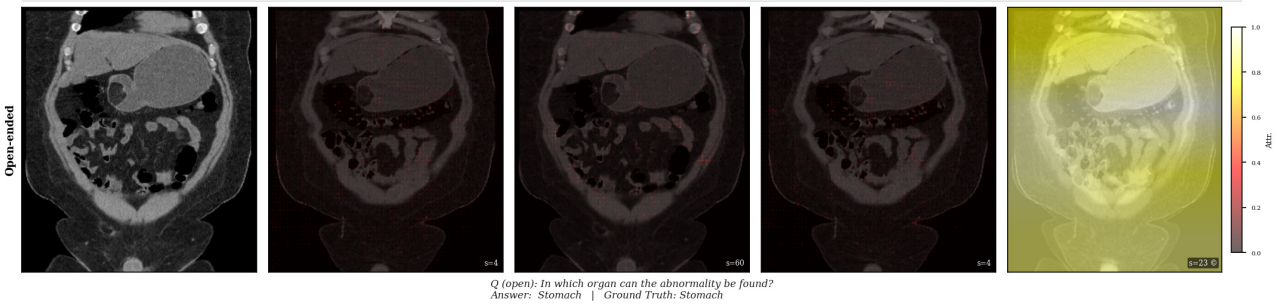


Figure 4.6: Open-ended attribution maps for abdominal CT case

Figure 4.6 focuses on the open-ended variant of the question. The model successfully generates the same answer without given choices. Accordingly, Integrated Gradients and Guided Grad-CAM preserve their activations tightly grouped around the gastric abnormality and surrounding interfaces. This slice and spatial invariance across prompting strategies highlights the robustness of input-level methods in visually anchoring abstract anatomical concepts, as opposed to the diffuse behavior of Layer Grad-CAM.

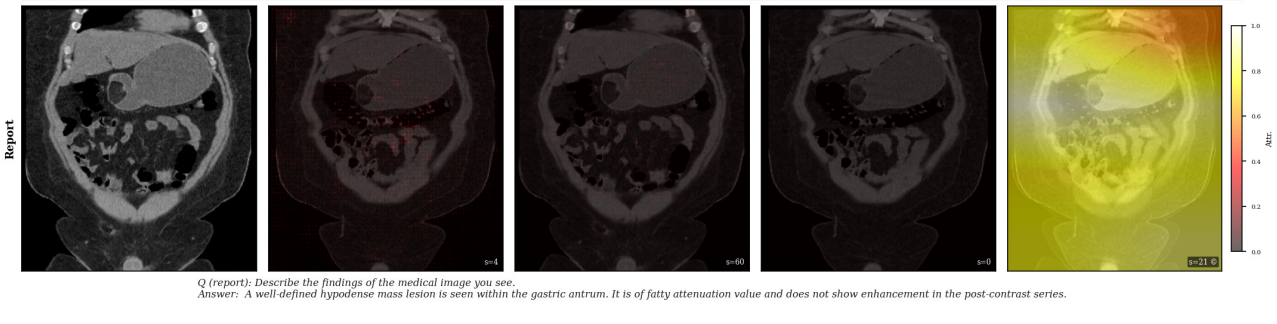


Figure 4.7: Report generation attribution maps for abdominal CT case

Figure 4.7 demonstrates the model's visual reasoning when tasked with producing a free-form diagnostic report of the findings. Compared to the constrained answers, the generated report addresses a broader spatial context of the stomach and adjacent organs. In response, input-level methods (notably Integrated Gradients and Saliency) correspondingly widen their localized pixel clusters to better encompass the whole diseased gastric landscape, proving that these methods structurally morph their attributions alongside the dynamically expanding diagnostic text.

4.3.1.3 Craniofacial CT Case Analysis

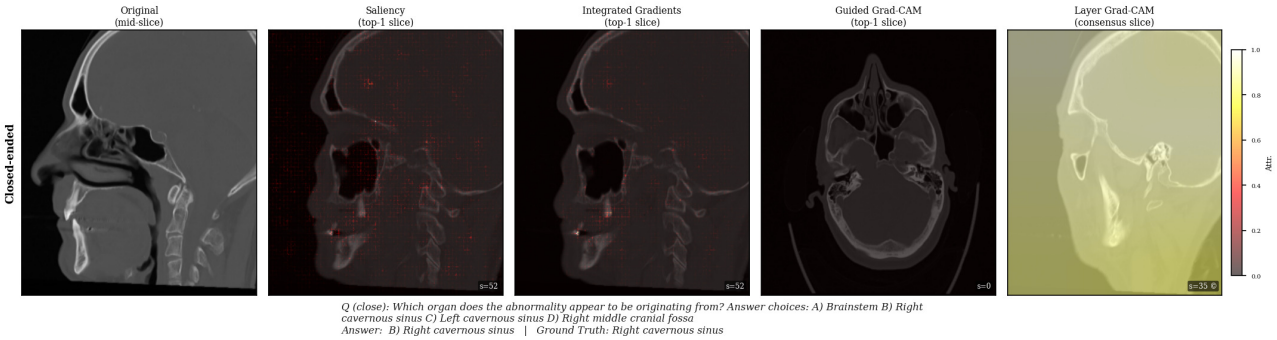


Figure 4.8: Closed-ended attribution maps for craniofacial CT case

Figure 4.8 details a craniofacial CT concerning a right cavernous sinus or skull-base abnormality with a closed-ended query. Highlighting complex facial bones and sphenoid structures, Saliency and Integrated Gradients precisely place fine red activations around the midface and posterior nasal regions. These localized maps identify anatomical features explicitly associated with the correct diagnosis. In contrast, Layer Grad-CAM produces a generalized thermal overlay covering the entire cranium without isolating the cavernous structures.

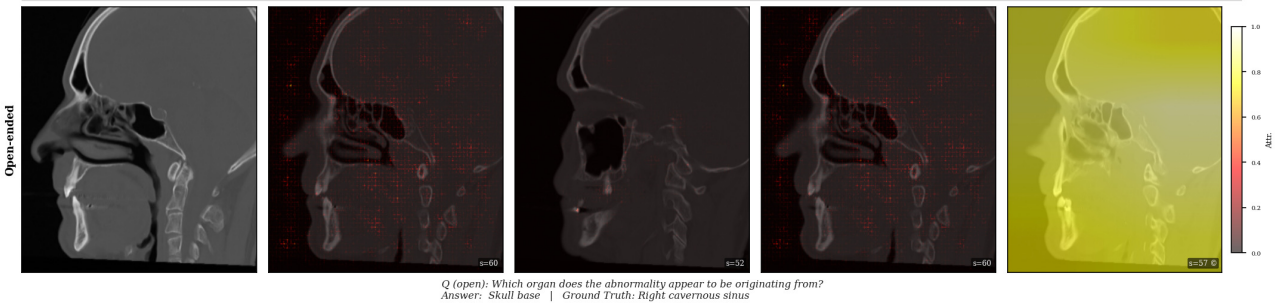


Figure 4.9: Open-ended attribution maps for craniofacial CT case

Figure 4.9 examines the open-ended query, exposing a distinct model ambiguity. The generated clinical text shifts its primary focus toward the "skull base." Functioning as an accurate diagnostic mirror, the integrated gradient and saliency attributions seamlessly shift and spread accordingly across the skull-base and facial-bone boundaries, demonstrating excellent question sensitivity. Layer Grad-CAM, conversely, remains diffuse and visually indifferent to the model's localized structural attention shift.

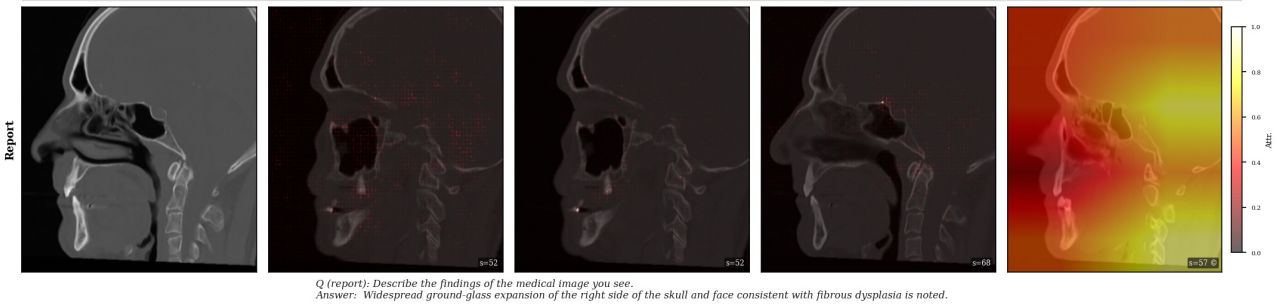


Figure 4.10: Report generation attribution maps for craniofacial CT case

Figure 4.10 shows the attribution maps aligned with the full report generation sequence. The model describes widespread ground-glass expansion covering the right skull and face. In direct response to the length and scope of the generated text, the input-space methods expand their localized clusters across the right craniofacial skeleton, validating that their relevance footprints proportionally adapt to the scale of the described pathology. Layer Grad-CAM fails to separate the cavernous sinus from the facial bone involvement, preserving a static glow.

Synthesizing these three anatomical evaluations reveals a consistent methodological hierarchy. Saliency captures sharp local gradients but often includes substantial speckle noise; Integrated Gradients suppresses more irrelevant background and provides the most stable lesion-adjacent localization; Guided Grad-CAM preserves high-frequency anatomical boundaries but can be affected by slice-selection mismatch; and Layer Grad-CAM repeatedly collapses the explanation into broad, low-resolution activation fields. These visual patterns reinforce the quantitative faithfulness analysis by showing that the most useful clinical explanations are those that remain spatially selective, question-sensitive, and anatomically aligned with the abnormality described in the answer.

4.3.2 Layer Grad-CAM Analysis

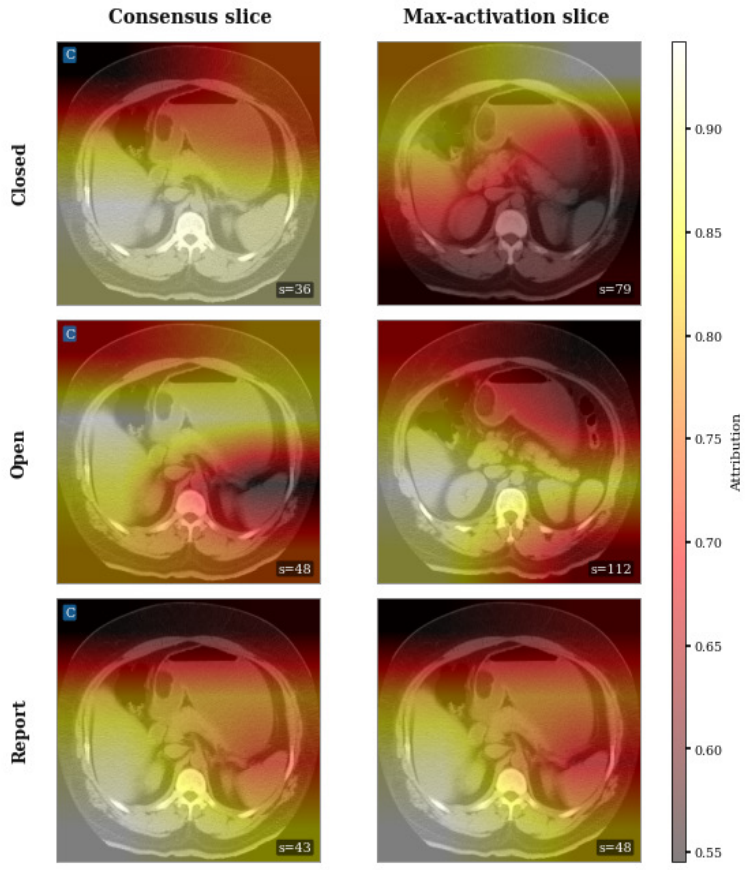


Figure 4.11: Comparison of Layer Grad-CAM at consensus vs. max-activation slices

Layer Grad-CAM produces a fundamentally distinct class of attribution map compared to the sparse, point-like clusters characteristic of input-space methods, as demonstrated in **Figure 4.11**. Instead of highly localized points, Layer Grad-CAM generates uniformly diffuse thermal heatmaps that blankets broad regional zones of the CT slice framework.

4.3.2.1 Consensus Slice Evaluation

When evaluated at the unified consensus slices (left column), Layer Grad-CAM exhibits the following behavior across the three task paradigms:

- **Closed-Ended VQA ($s = 36$):** The method produces a broad, sprawling heatmap covering the vast majority of the abdominal cross-section, failing to isolate or concentrate score mass on any specific diagnostic anatomy.
- **Open-Ended VQA ($s = 48$):** The resulting heatmap remains similarly diffuse, demonstrating no meaningful structural mutations or spatial adjustments in response to the unconstrained text generation path.
- **Report Generation ($s = 43$):** The attribution pattern is once again broadly dispersed across the slice, showing no discernible shift in spatial focus toward the clinically relevant splenic pathology.

Across all three question variants at the consensus depth, Layer Grad-CAM fails to isolate the localized anatomically or semantically relevant diagnostic regions. This pronounced absence of task-driven question sensitivity stands in stark contrast to the input-space frameworks.

The underlying mathematical cause for this diffuseness stems from the coarse spatial resolution ($D' \ll D, H' \ll H, W' \ll W$) of the final target encoder layer (`encoder.s4.1.dwconv.c1`), compounded by the global average pooling operation used to derive channel importance weights, which fundamentally strips out fine-grained spatial gradients.

4.3.2.2 Consensus vs. Maximum-Activation Slice Discrepancy

A direct comparison between Layer Grad-CAM's performance at the consensus slice versus its own localized maximum-attribution depth (right column) further highlights these architectural limitations.

When the algorithm is allowed to select its own peak activation slice, the depth coordinates shift abruptly to completely distinct and anatomically non-selective depths (e.g., migrating to $s = 79$ and $s = 112$ instead of aligning with the consensus markers at $s = 36$ and $s = 48$). At these individual peak slices, the method generates completely uniform, washed-out heatmaps that cover the entire scan matrix indiscriminately, regardless of the prompt format or clinical context under evaluation.

4.4 Quantitative Faithfulness Evaluation

4.4.1 Voxel Deletion Results

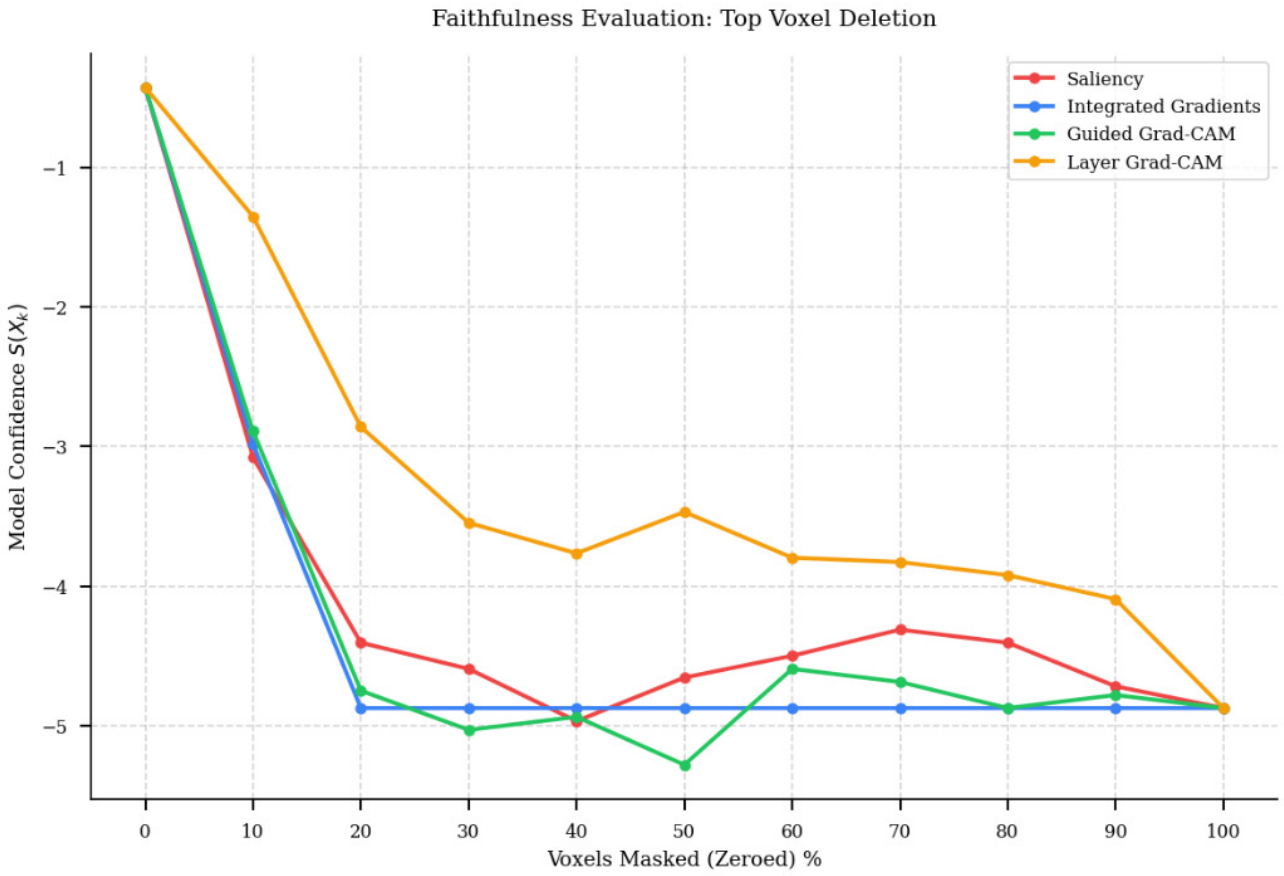


Figure 4.12: Voxel deletion curves for all attribution methods

The voxel deletion protocol is applied to all four attribution methods on the same input volumes (Figure 4.12). For each method, the top-k most important voxels are progressively masked, and the resulting scalar objective S is re-evaluated at each deletion step.

The quantitative results reveal a clear and consistent hierarchy among the four methods:

- **Integrated Gradients** achieves the steepest and most monotonic confidence drop among all four methods. Its deletion curve reaches the score floor by approximately 20% of voxels masked. This rapid saturation confirms that the compact voxel clusters highlighted by Integrated Gradients are causally necessary for the model's response.
- **Saliency** follows closely behind Integrated Gradients. Its deletion curve also saturates near the score floor at approximately 20% of voxels masked, though with slightly more oscillation in the intermediate deletion steps.
- **Guided Grad-CAM** produces a deletion curve that is broadly comparable to those of

Saliency and Integrated Gradients, also saturating near the score floor at approximately 20% deletion.

- **Layer Grad-CAM** degrades substantially slower than the three other methods. At 30% deletion, its confidence remains far above the score floor reached by the input-space methods. This slow degradation reveals that the broad activations highlighted by Layer Grad-CAM are dominated by anatomically prominent but causally irrelevant voxels.

Interpretation of the Faithfulness Hierarchy The faithfulness hierarchy observed in the voxel deletion results (Integrated Gradients $>$ Saliency $\approx$ Guided Grad-CAM $\gg$ Layer Grad-CAM) has several important implications.

The superior performance of Integrated Gradients can be attributed to its axiomatic design. The completeness axiom guarantees that the total attribution exactly equals the difference in model output between the actual input and the baseline, ensuring that no attribution "leaks" or is lost during the computation.

The comparable performance of Saliency and Guided Grad-CAM demonstrates that both methods, despite their different computational approaches, successfully identify the causally relevant voxels. The fact that both methods reach the score floor at approximately the same deletion threshold confirms that the input-space gradient signal provides a reliable indicator of causal importance.

The poor performance of Layer Grad-CAM quantitatively confirms that feature-level activation maps are an unreliable indicator of causal importance in multimodal generative architectures. The coarse spatial resolution of the target encoder layer, combined with the global average pooling operation, produces attribution maps that are too spatially diffuse to identify the specific voxels that drive the model's prediction.

4.5 Discussion

4.5.1 Input-Level Methods Outperform Feature-Level Methods

The most significant finding of this study is the clear superiority of input-level attribution methods over feature-level methods in the context of generative volumetric medical VLMs. Both the qualitative analysis (which demonstrates the spatial selectivity and question sensitivity of input-space methods) and the quantitative evaluation (which confirms their causal faithfulness through the voxel deletion protocol) consistently support this conclusion.

This finding has practical implications for the deployment of XAI methods in clinical set-

tings. When a clinician needs to understand why a medical VLM generated a particular report or answered a clinical question in a specific way, input-level methods such as Integrated Gradients and Saliency provide more trustworthy and informative explanations than feature-level methods such as Layer Grad-CAM.

4.5.2 Layer Grad-CAM's Limitations in Multimodal Architectures

The poor performance of Layer Grad-CAM in this study does not necessarily imply that the method is inherently flawed. Layer Grad-CAM has been shown to be effective for explaining discriminative models where the target layer directly precedes the classification head. However, in the multimodal VLM architecture used in this study, the target layer is located in the visual encoder, which is separated from the final decision-making process by a multimodal projector and an entire language model decoder. The gradient signal must propagate backward through all of these additional components, introducing opportunities for gradient attenuation and diffusion.

Furthermore, the global average pooling operation used to compute channel importance weights collapses all spatial information within each channel into a single scalar weight, making it impossible for Layer Grad-CAM to achieve the voxel-level precision needed for clinically meaningful localization.

4.5.3 Question Sensitivity as Evidence of Meaningful Attribution

The observation that all three input-space methods produce question-sensitive attribution maps is a significant finding. The attribution maps shift their spatial focus when the clinical question type changes, providing evidence that the model is genuinely reasoning about the relationship between the visual content and the specific clinical query. The observed shifts, particularly the migration toward cranial slices containing the spleen during report generation (consistent with the splenic pathology expressed in the generated text), strongly suggest that the model's attention is driven by the clinical relevance of each region to the specific task.

This question sensitivity is especially important for clinical deployment, as it means that the explanations are not generic but are tailored to the specific clinical context.

4.5.4 The Faithfulness-Resolution Trade-off

The results reveal a trade-off between the spatial resolution of an attribution method and its computational cost. Integrated Gradients, which produces the most faithful and spatially precise attribution maps, requires 50 forward and backward passes through the model, making

it approximately 50 times more computationally expensive than Input-Level Saliency. Saliency offers a compelling alternative for time-sensitive clinical workflows: it produces reasonably faithful attribution maps at a fraction of the computational cost [33, 32].

4.6 Conclusion

This chapter has presented the full implementation and experimental evaluation of the proposed XAI attribution framework for generative volumetric medical vision-language models. The implementation details confirmed the feasibility of applying gradient-based attribution methods to the frozen Med3DVLM model through the two-pass teacher forcing pipeline, within the computational constraints of a Kaggle environment equipped with an NVIDIA P100 GPU.

The qualitative results demonstrated that input-level methods, particularly Integrated Gradients, produce sparse and spatially selective attribution maps that shift meaningfully with question type, confirming that the attributions are semantically grounded in the model's clinical reasoning. Layer Grad-CAM, by contrast, produced uniformly diffuse heatmaps across all question types, revealing a fundamental limitation of feature-level methods in multimodal volumetric architectures. These qualitative findings were corroborated by the quantitative voxel deletion evaluation, which confirmed that input-level methods achieve a significantly steeper and more monotonic confidence drop than feature-level methods, establishing Integrated Gradients as the most causally faithful attribution method in this setting.

Taken together, these results demonstrate that the proposed attribution framework produces meaningful, spatially precise, and causally faithful explanations for generative volumetric medical VLMs, advancing the state of the art in explainable medical AI beyond the 2D unimodal settings that dominate the existing literature. The following chapter presents RadEx, the clinical platform developed to demonstrate the practical applicability of these findings in a real-world medical imaging context.

Chapter 5

Platform Implementation

5.1 Introduction

The preceding chapters established the theoretical foundations of explainable artificial intelligence in healthcare (Chapter 2), presented a methodology for applying gradient-based attribution methods to generative volumetric vision-language models (Chapter 3), and reported experimental results demonstrating the superiority of input-level attribution methods in terms of spatial precision and causal faithfulness (Chapter 4). While these contributions advance the scientific understanding of post-hoc explainability in multimodal medical AI, their practical impact depends on the availability of software systems that can deliver such explanations to clinicians within real-world diagnostic workflows.

This chapter presents RadEx, a web-based clinical platform developed as part of this work to bridge the gap between the theoretical attribution framework and its practical deployment. RadEx integrates multimodal diagnostic inference with visual explanation overlays, enabling clinicians to upload medical imaging studies, formulate clinical queries through natural language, and receive AI-generated diagnostic outputs accompanied by spatially grounded attribution maps. The platform is designed to embody the principle that explainability is not an auxiliary feature but a fundamental requirement for the responsible deployment of AI in clinical practice.

The remainder of this chapter is organized as follows. Section 5.2 presents the public-facing interface through representative screenshots that introduce the platform's visual identity and key capabilities. Section 5.3 describes the system architecture, detailing the three-tier design that separates the client interface, the API layer, and the inference backend. Section 5.4 presents the clinical workflow through which users interact with the platform. Section 5.5 describes the integration of attribution-based explanation overlays into the medical image viewer. Section 5.6 details the dashboard interface and its navigation structure. Section 5.7 introduces the educational extension, RadEx Edu, designed for residency training. Finally, Section 5.8 concludes with a summary of the platform's contributions and its relationship to the experimental work presented in earlier chapters.

5.2 Platform Overview

Before describing the technical architecture in detail, this section introduces the RadEx platform through its public-facing interface, providing the reader with a visual overview of the system's design and capabilities.

Figure 5.1 presents the landing page of the RadEx platform. The header communicates the system's core objective: enabling clinicians to inspect the reasoning of AI diagnostic models before relying on their outputs in clinical practice. The Clinical Alignment Viewer preview juxtaposes raw anatomical slices alongside color-coded attribution maps, displays the VQA prompt and model summary, and identifies the active attribution method (Layer Grad-CAM), illustrating the multi-tab explainability controls available within the authenticated dashboard.

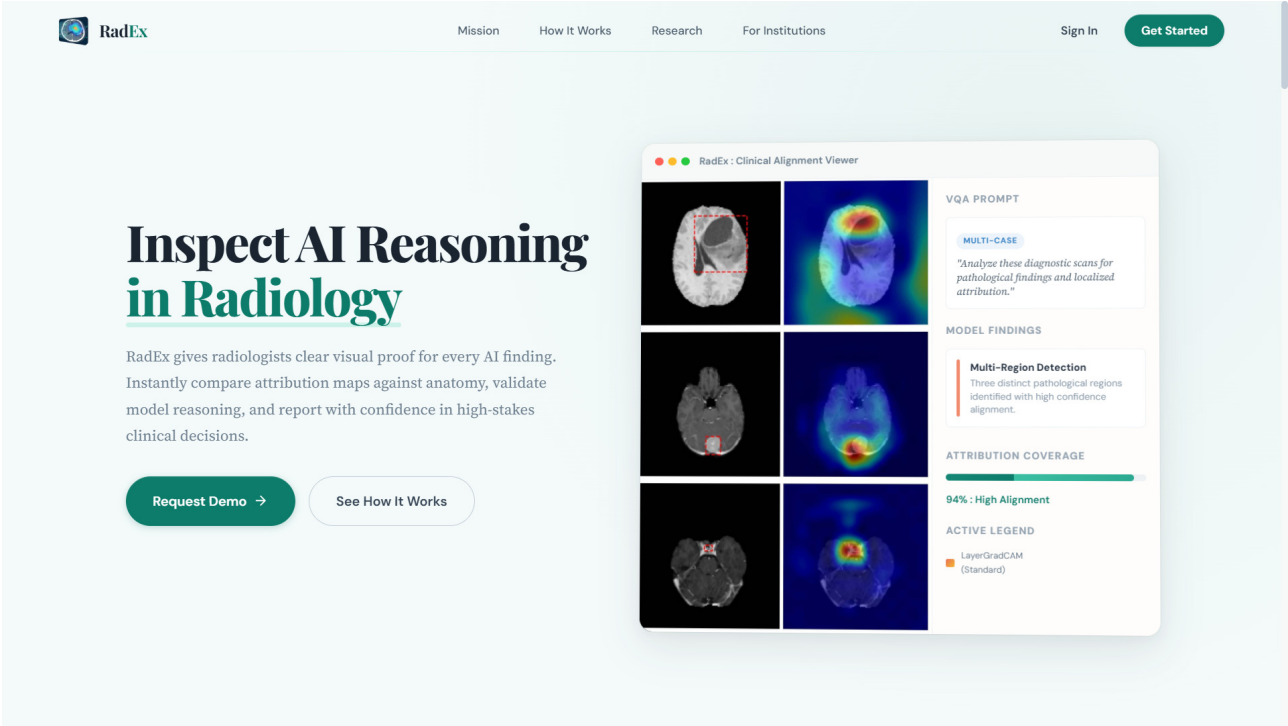


Figure 5.1: RadEx landing page and Clinical Alignment Viewer.

Figure 5.2 presents the volumetric analysis section of the public page. This section demonstrates multi-planar attribution coverage across axial, sagittal, and coronal views. The Volumetric Attribution Overlay panel illustrates a closed-ended VQA task: a multiple-choice question about which organ harbors a mass, the selected answer, and its agreement with a reference label. Coronal images pair a Grad-CAM-style overlay with the unprocessed slice, allowing the viewer to inspect the spatial regions that the model identifies as diagnostically relevant.

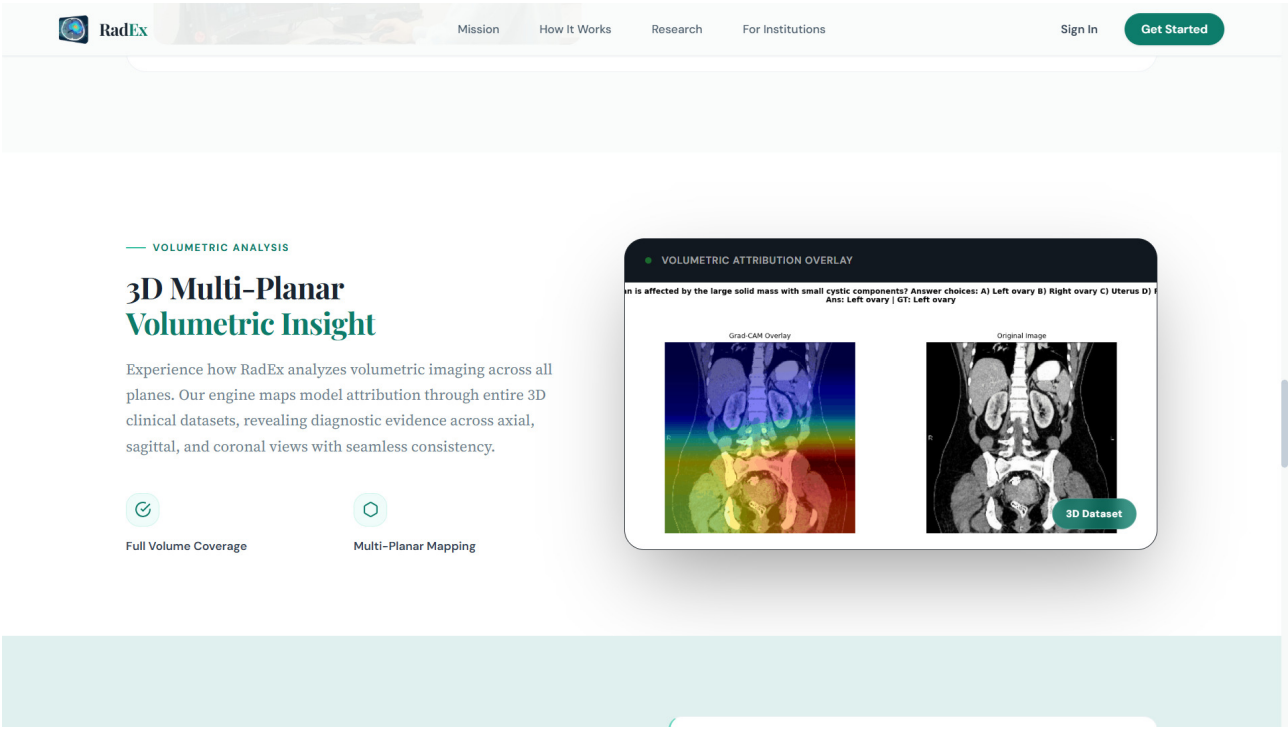


Figure 5.2: RadEx volumetric attribution overlay demonstration.

5.3 Platform Architecture

RadEx is implemented as a browser-based application that combines multimodal diagnostic inference with authentic gradient-based attribution explanation overlays for volumetric medical image review. The system architecture follows a structured three-tier design that cleanly separates presentation, application logic, and heavy deep-learning inference concerns, as described below.

5.3.1 Frontend Layer

The client application is built using React 19 with TypeScript and bundled through Vite 8. Plain CSS is used for styling, with BEM-style class naming conventions (e.g., `dash__sidebar`, `hero__mockup`) to enforce encapsulation and prevent style conflicts between views. React Router v7 manages navigation across six routes: a public landing page (`HomePage`), three authentication views (`LoginPage`, `RegisterPage`, `ResetPasswordPage`), the primary clinical workspace (`DashboardPage`), and an academic demonstration view (`UniversityDashboardPage`).

Shared interface components, including the navigation bar (`Navbar`) and footer (`Footer`), are rendered consistently across all routes. Layouts are optimized for high-resolution desktop and tablet display widths typical of hospital radiology reading rooms and clinical workstations.

5.3.2 API Layer

The backend is implemented as an Express.js 5 server that exposes RESTful endpoints for volumetric image upload, report generation, and diagnostic record retrieval. The browser communicates with the server via standard `fetch` requests using `FormData` for multipart file uploads. The `multer` middleware handles file storage, writing uploaded scans to disk under unique timestamped filenames.

Two primary endpoints define the API surface: `POST /api/generate-report`, which accepts a volumetric file, a task type identifier, and a clinical prompt and `GET /api/reports`, which returns a list of previously generated reports. Report payloads and associated metadata are serialized as JSON files alongside the stored images, enabling stateless server operation without requiring server-side session management. Cross-Origin Resource Sharing (CORS) is configured for the development origin to support local deployment.

The server natively inspects the uploaded file to ensure it matches supported volumetric medical formats, including NIfTI, NRRD, MHA, and DICOM [49, 50, 51, 52]. Once validated, the server routes the data directly into the core volumetric processing pipeline described below.

5.3.3 Inference Layer: Volumetric Data Pipeline

The core processing tier routes incoming volumetric medical scans to a remote, high-performance Kaggle-hosted inference service running the Med3DVLM model with the dual-pass teacher-forced attribution framework.

Remote Inference Coordination The Express.js backend relays the validated volumetric scan file and the corresponding clinical prompt to the remote Kaggle service endpoint via an HTTP POST request using `axios` and `FormData`, with the connection established through a secure ngrok tunnel. The remote Kaggle service loads the volumetric scan, feeds it through the frozen Med3DVLM architecture, and sequentially executes the four mathematically grounded gradient-based attribution methods: Input-Level Saliency, Integrated Gradients, Layer Grad-CAM, and Guided Grad-CAM.

XAI Overlay Generation and Fusing The Kaggle endpoint computes the raw mathematical gradients and extracts the corresponding relevance values across the 3D volume grid. The service identifies the peak activation slices for each XAI technique and pre-renders the resulting 3D relevance maps directly onto the respective grayscale structural slices as colorful transparency overlays.

The service returns a structured JSON payload to the main server containing:

- The model's complete generated text answer sequence $\hat{A}$.
- Per-method attribution metadata, including the determined consensus slice indices and operational parameters.
- Base64-encoded overlay images where the 3D relevance maps $\mathbf{M}$ are fused onto the source slices.

These overlay images are decoded by the API server and written to the local disk file system as PNG assets, making them instantly available for synchronized client-side display within the custom Explainability Viewer layout.

This single, unified pipeline guarantees that every diagnostic report generated on the platform is accompanied by authentic, path-faithful gradient attributions computed on the actual underlying model, ensuring rigorous transparency for clinical decision support.

5.4 Clinical Workflow

The platform guides the user through a sequential four-stage workflow that mirrors the structure of standard radiology tasks: image ingestion, analysis configuration, clinical query formulation, and inference execution with result review.

1. **Image Upload.** The user uploads a medical imaging study through a drag-and-drop interface that accepts DICOM, JPEG, PNG, and NIfTI (`.nii.gz`) file formats, with a maximum file size of 500 MB. The interface confirms the selected file before transmitting it to the server.
2. **Analysis Configuration.** Two configuration selectors allow the user to specify the parameters of the analysis. The first selector specifies the imaging modality (MRI/CT or X-ray), and the second specifies the VQA task type (closed-ended question answering, open-ended question answering, or structured report generation). Contextual help text accompanies each option to guide selection.
3. **Clinical Query Formulation.** A text input field captures the clinician's natural-language question or reporting instruction. Placeholder text adapts dynamically to the selected task type, providing examples appropriate to binary screening questions, free-form clinical descriptions, or comprehensive report requests.

4. **Inference Execution and Result Delivery.** Upon initiating the analysis, the client constructs a `FormData` payload containing the image, task type, and clinical prompt, and submits it to the `/api/generate-report` endpoint. A loading indicator is displayed during processing. When the server returns the structured JSON response, the client navigates to the Overview panel, populates the Explainability Viewer with the generated overlays, and appends corresponding entries to the Analyses and Reports views to maintain interface consistency.

Figure 5.3 illustrates the upload interface with a NIfTI volume selected, the MRI/CT modality and report-generation task type configured, a clinical prompt entered, and the analysis execution control visible.

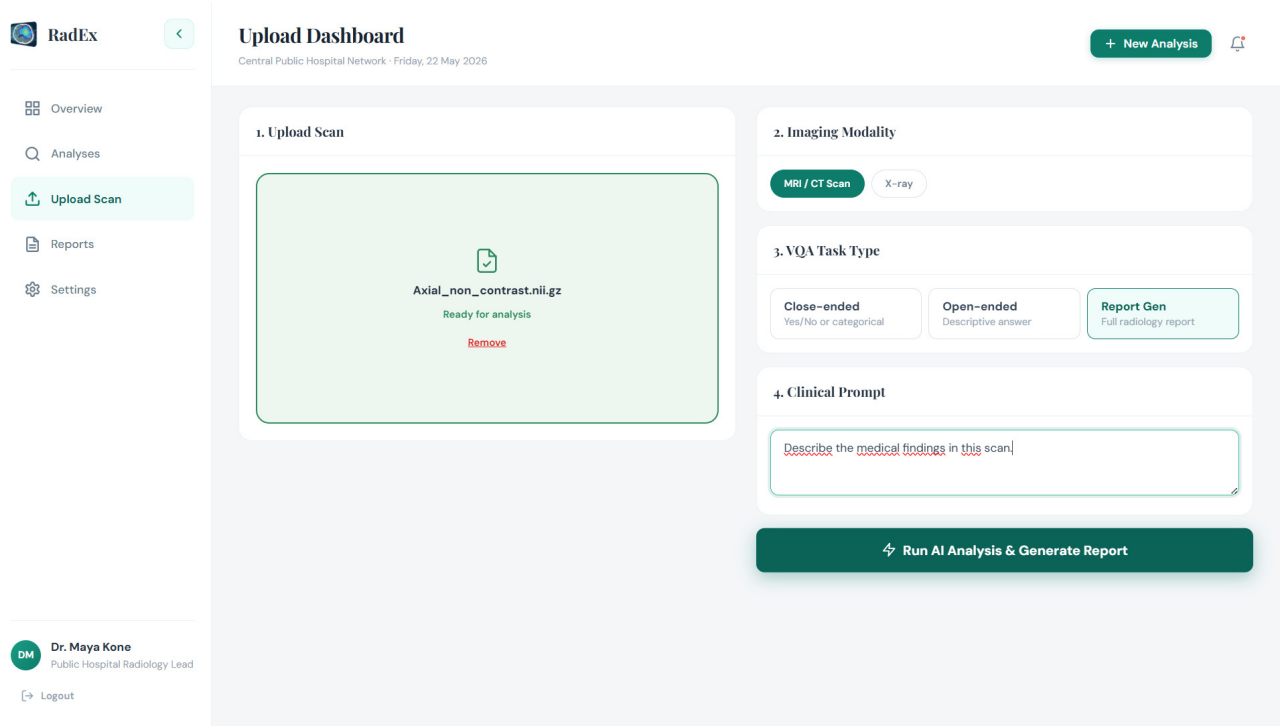


Figure 5.3: RadEx scan upload and analysis configuration interface.

5.5 Explainability Integration

A central design principle of RadEx is the tight integration of attribution-based explanations into the medical image viewer. Rather than presenting explanation maps as supplementary outputs accessible only through separate interfaces, the platform embeds them directly into the primary diagnostic review environment. This design reflects the dissertation's broader argument that explainability must be an integral component of clinical AI systems rather than an optional addition.

5.5.1 The Explainability Viewer

The Explainability Viewer provides a tabbed interface that maps API-returned explanation fields onto the displayed medical image. Five tabs control the visualization canvas, each associated with a distinct color accent for immediate visual differentiation:

- **Original Slice:** Displays the unmodified base imaging slice, providing a clear anatomical reference for comparison with the generated attribution overlays.
- **Layer Grad-CAM:** Renders a Grad-CAM-style composite overlay that highlights the coarse anatomical regions associated with the model's response, corresponding to the feature-level attribution method described in Chapter 3.
- **Saliency:** Displays a saliency-style map emphasizing locations where small input perturbations would most strongly affect the model's output confidence.
- **Integrated Gradients:** Presents the integrated-gradient attribution visualization, corresponding to the axiomatic attribution method that accumulates gradients along an interpolation path from baseline to input.
- **Guided Grad-CAM:** Renders a guided Grad-CAM composite that combines the spatial localization of Grad-CAM with the fine-grained edge structure of guided backpropagation.

The viewer sidebar presents case metadata including patient age, the submitted VQA prompt, the principal diagnostic finding, a confidence score indicator, and the number of explanation overlays generated. A navigation control provides direct access to the Reports view for structured review and export.

5.5.2 Attribution Maps in Structured Reports

The Reports view extends the explanation presentation beyond the interactive viewer by embedding the four attribution map visualizations within a structured clinical narrative. The JSON response from the inference layer is expanded into standard report sections (patient demographics, clinical history, findings, diagnostic impressions, and recommendations), and the four corresponding heatmap images are arranged in a single comparative gallery. This design enables reviewers to contrast the spatial emphasis patterns of different attribution methods within the context of the full clinical narrative without returning to the interactive viewer.

5.6 Dashboard Interface and Navigation

This section describes the organizational structure and principal views of the RadEx dashboard, through which authenticated users access all clinical and analytical functionality.

5.6.1 Navigation Structure

The authenticated application employs a persistent left navigation rail with five primary destinations: Overview, Analyses, Upload Scan, Reports, and Settings. Each destination is represented by a labeled button with an inline SVG icon. On viewports narrower than approximately 900 pixels, the navigation rail collapses into a hamburger-triggered drawer with a dimmed backdrop overlay, ensuring usability across standard clinical workstation configurations.

5.6.2 Overview

The Overview panel serves as the primary landing view upon authentication. It combines aggregate performance metrics (total analysis count, heatmaps rendered, mean confidence score, and mean inference latency) with the Explainability Viewer for the most recently analyzed case. Supplementary visualizations include a weekly analysis histogram and a horizontal bar chart of attribution method usage frequency. Figure 5.4 illustrates the Overview panel with a brain MRI displaying a Grad-CAM overlay in the central viewer, tabs for the remaining attribution methods, case metadata and confidence score in the right column, and usage analytics in the peripheral panels.

5.6.3 Reports

The Reports view transforms the structured JSON response into a clinical narrative organized into standard sections: primary diagnostic impression, extended findings, referral recommendations (including specialty, urgency, clinical justification, and recommended actions), and a four-panel attribution map gallery. Figure 5.5 presents the report layout with Layer Grad-CAM, Saliency, Integrated Gradients, and Guided Grad-CAM visualizations displayed side by side on the same anatomical slice, enabling direct visual comparison of the spatial emphasis patterns produced by each attribution method.

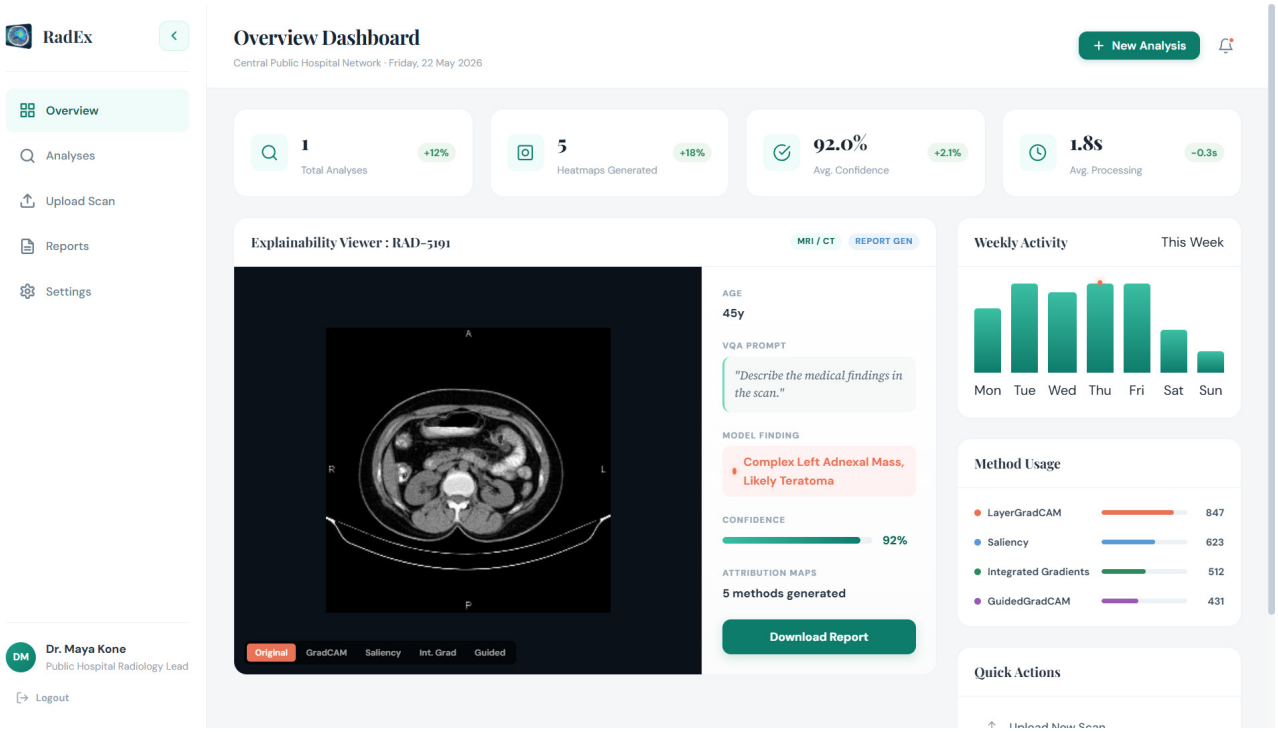


Figure 5.4: RadEx Overview dashboard and Explainability Viewer.

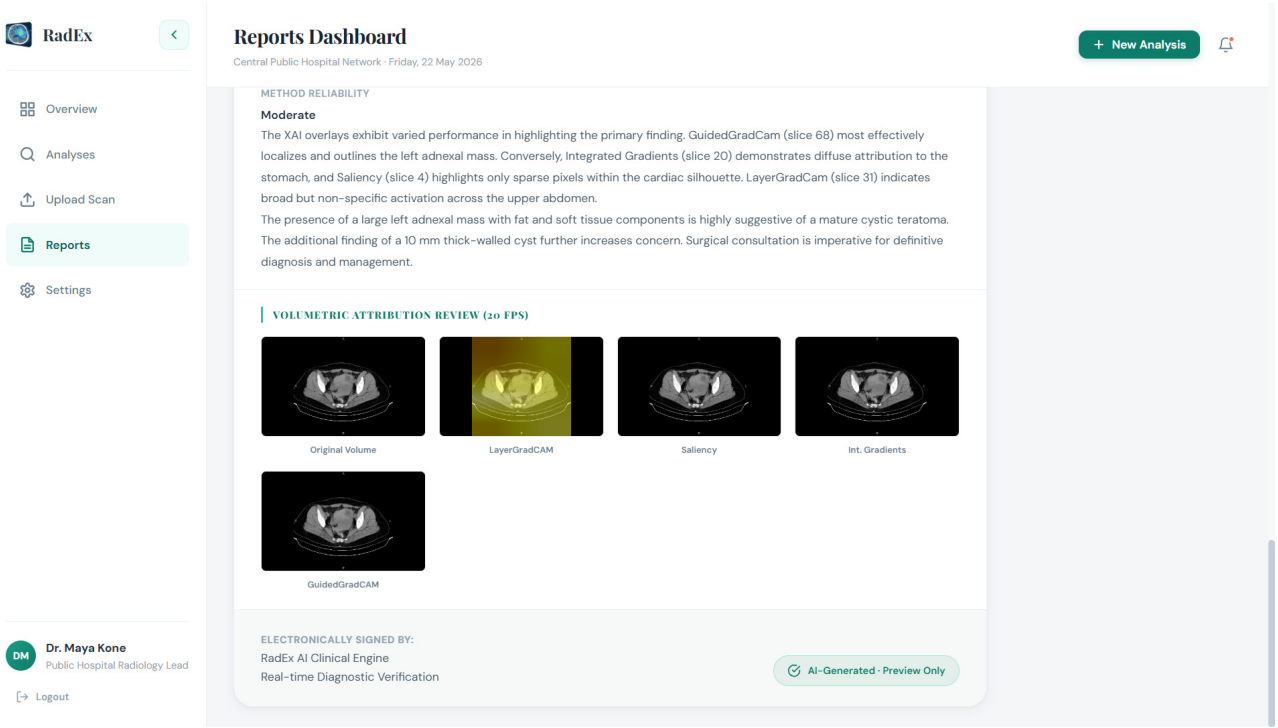


Figure 5.5: RadEx structured report with attribution map gallery.

5.6.4 Analyses

The Analyses view provides a searchable data table of all processed cases. Full-text search is supported on case identifiers and finding text, and status filter chips (All, Reviewed, Pending, Flagged) enable rapid case triage. Table columns display case identifiers, patient age, imaging modality, task type, condensed finding text, model confidence score, processing date, and review status. Selecting a row loads the corresponding study in the Explainability Viewer for detailed review. Figure 5.6 illustrates this layout, including summary counts in the table footer.

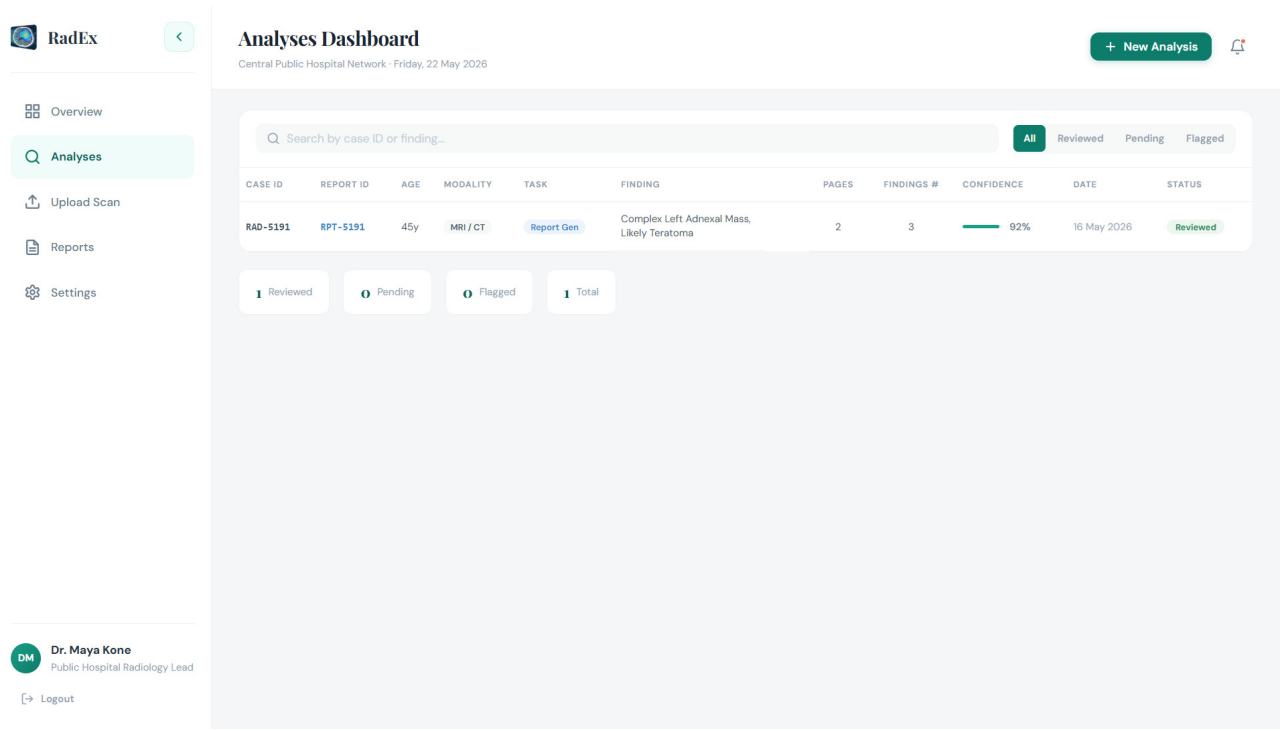


Figure 5.6: RadEx Analyses dashboard and case table.

5.7 Academic Platform: RadEx Edu

To extend the utility of explainable medical AI into the domain of clinical education, the platform includes a specialized educational module titled RadEx Edu. This extension is designed as a training laboratory for radiology residents, allowing them to practice diagnostic reporting on a curated library of complex cases while interacting with the AI-driven explainability framework.

5.7.1 Resident Progress and Milestones

Figure 5.7 illustrates the resident's performance dashboard. The interface provides a quantitative metric known as RadEx Report Alignment, which quantifies how closely the resident's

findings match the established ground truth across completed modules. Academic milestones and a detailed breakdown of case-specific scores provide residents with a clear roadmap for their professional development within the training program.

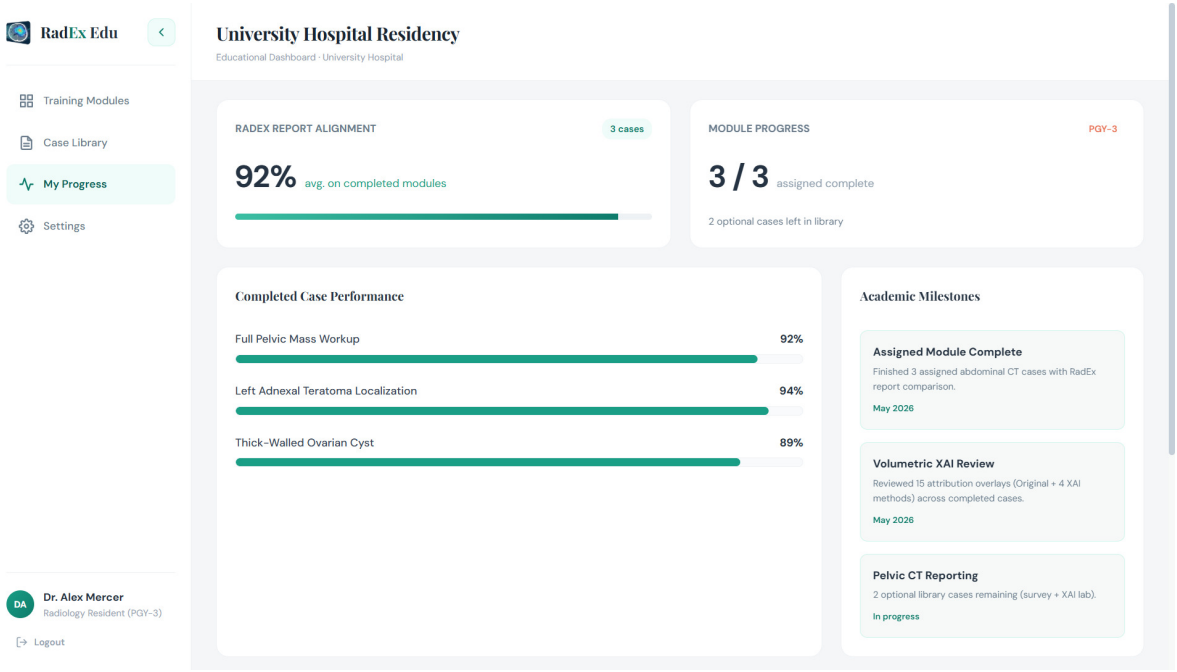


Figure 5.7: RadEx Edu resident progress tracking dashboard.

5.7.2 Case Modules and Library

The training curriculum is organized into structured modules, as shown in Figure 5.8. Each module contains a set of assigned practice cases characterized by their anatomical focus and difficulty level. The inclusion of "Volumetric XAI" tags indicates cases where the full suite of gradient-based attribution methods is available for review, encouraging residents to explore the model's reasoning in depth.

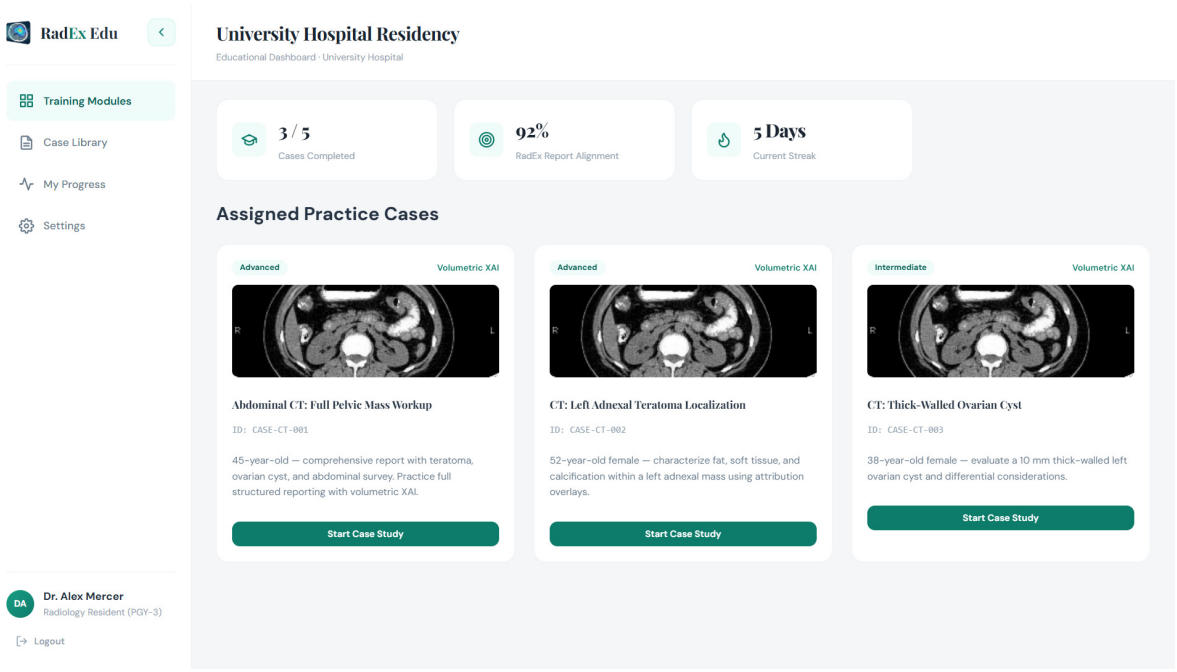


Figure 5.8: RadEx Edu training modules and assigned practice cases.

For self-directed study, residents can access a comprehensive Case Library (Figure 5.9). This searchable archive includes filters for imaging modality and diagnostic difficulty, enabling residents to hone their skills on specific pathologies or rare clinical presentations.

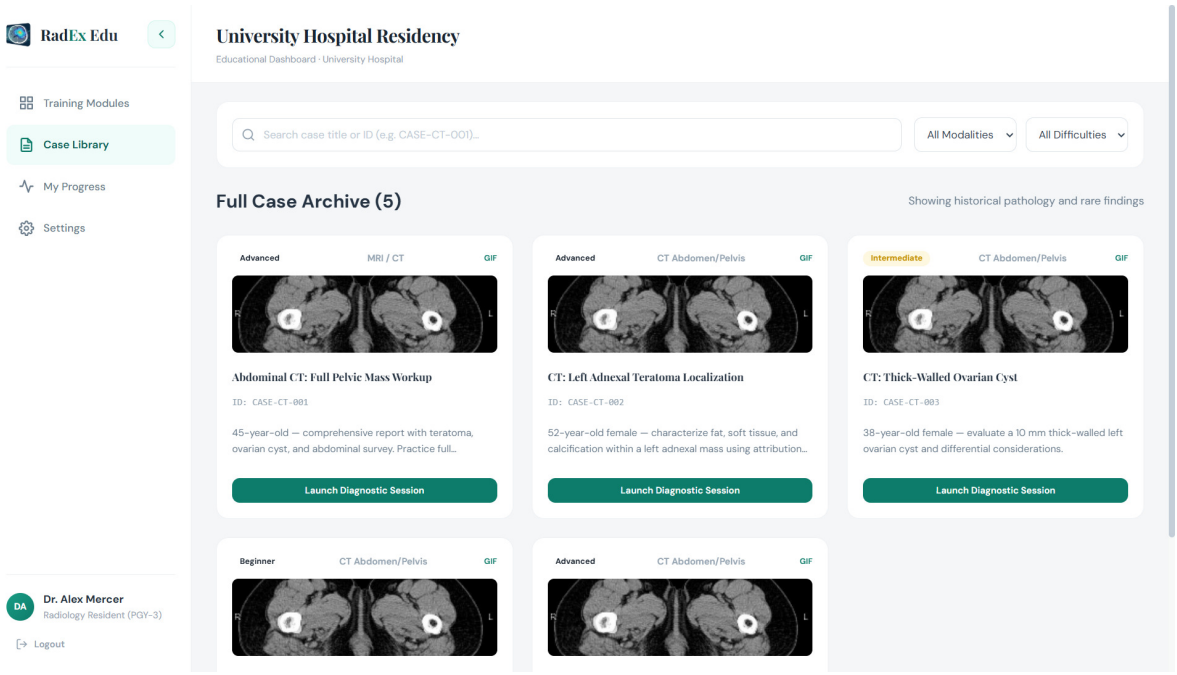


Figure 5.9: RadEx Edu searchable clinical case archive.

5.7.3 Interactive Reporting Practice

The core educational experience is centered on the Diagnostic Image Viewer (Figure 5.10). Within this interface, residents perform a structured workup of the volumetric scan, entering their findings and diagnostic impressions in real-time. The viewer permits active toggling between different attribution overlays (e.g., Grad-CAM, Saliency), allowing the resident to visually corroborate their own anatomical observations with the model's spatial relevance maps.

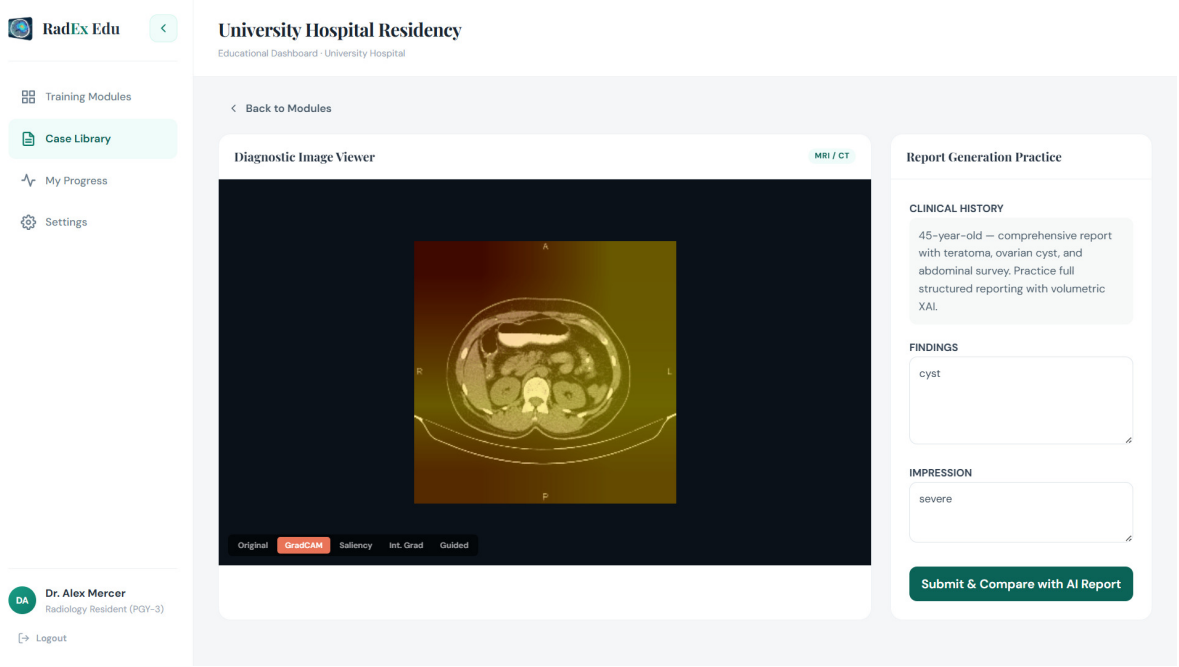


Figure 5.10: RadEx Edu diagnostic image viewer and reporting interface.

Upon submission, the platform provides an immediate comparison between the resident's report and the "RadEx AI Ground Truth" (Figure 5.11). This comparison highlights exactly which findings were correctly identified and which clinically significant features may have been overlooked, providing a powerful, evidence-based feedback mechanism for radiological training.

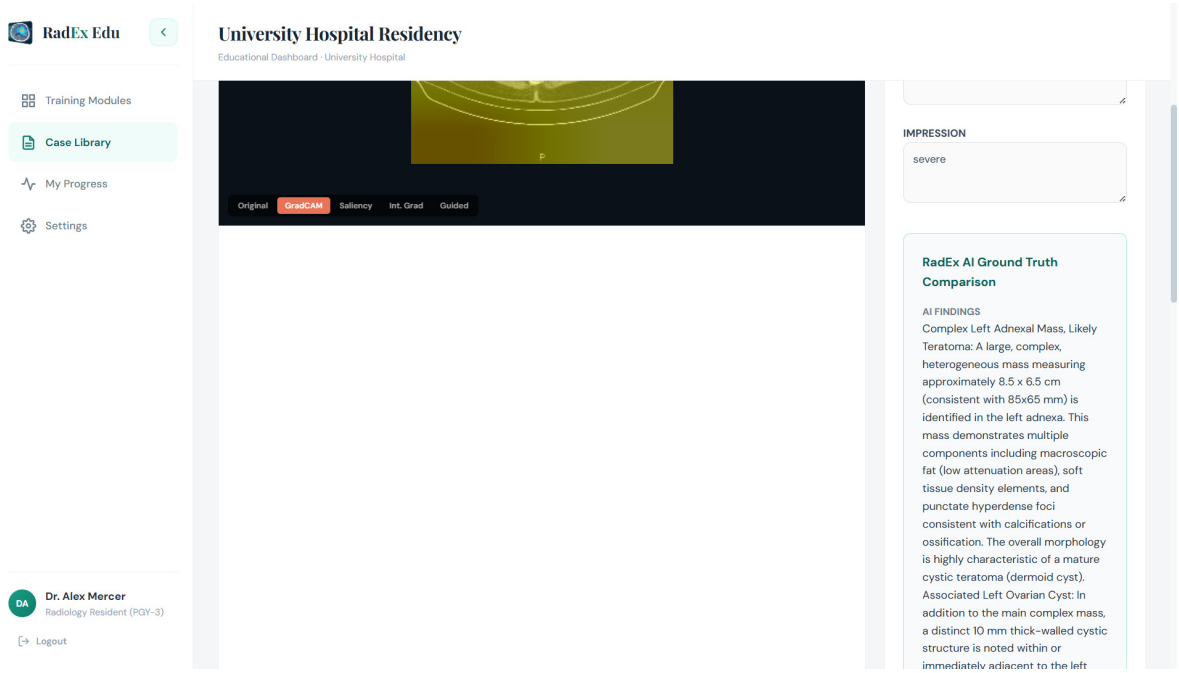


Figure 5.11: RadEx AI ground truth comparison for resident feedback.

5.8 Conclusion

This chapter has presented RadEx, a web-based clinical platform that integrates multimodal diagnostic inference with attribution-based visual explanations within a unified radiology review environment. The system architecture separates concerns across three tiers: a React-based client application, an Express.js API layer, and compositing backend, supplemented by a FastAPI service for volumetric NIfTI data browsing.

The clinical workflow follows a sequential four-stage process of image upload, analysis configuration, clinical query formulation, and inference execution, delivering results through an Explainability Viewer that embeds four attribution method overlays directly into the diagnostic review interface and a Reports view that presents structured clinical narratives alongside comparative attribution map galleries. The dashboard architecture provides authenticated users with aggregate performance analytics, case management capabilities, and configurable preferences.

The platform demonstrates how the explainability principles investigated in this dissertation can be translated from experimental evaluation into a deployable clinical tool. Within this scope, RadEx illustrates that XAI-oriented interface design can be integrated as a fundamental component of multimodal radiology assistance systems, supporting the broader goal of responsible and transparent clinical AI deployment.

General Conclusion

This dissertation has investigated the application of post-hoc gradient-based attribution methods to generative volumetric medical vision-language models, addressing a significant and largely unexplored gap at the intersection of explainable AI and multimodal medical imaging. By combining rigorous scientific evaluation with a practical clinical platform, this work makes contributions that are simultaneously theoretical, methodological, and applied.

The central methodological contribution of this dissertation is the development of a differentiable attribution framework that renders the stochastic autoregressive generation process of Med3DVLM amenable to gradient-based attribution. By introducing a teacher forcing strategy and defining a scalar attribution target based on the cumulative log-likelihood of the model's generated response, the proposed framework converts the non-differentiable generation path into a fully deterministic and end-to-end differentiable computation graph. This technical contribution is a prerequisite for any gradient-based XAI method to be applied to generative models, and its absence in the existing literature represented one of the key gaps that motivated this work.

Building on this framework, four gradient-based attribution methods spanning input-level and feature-level families were systematically applied and evaluated across three clinical question types. The qualitative analysis demonstrated that input-level methods, particularly Integrated Gradients, produce sparse and spatially selective attribution maps that shift meaningfully with question type, confirming that the attributions are semantically grounded in the model's clinical reasoning rather than being arbitrary visual artifacts. Layer Grad-CAM, by contrast, produced uniformly diffuse heatmaps across all question types, revealing a fundamental limitation of feature-level methods when applied to multimodal volumetric architectures. These qualitative findings were corroborated by a quantitative voxel deletion faithfulness evaluation, which confirmed that input-level methods achieve a significantly steeper and more monotonic confidence drop than feature-level methods, establishing Integrated Gradients as the most causally faithful attribution method in this setting.

The practical contribution of this dissertation is RadEx AI, a web-based clinical platform that integrates the proposed attribution framework into a specialized medical imaging interface. The platform allows clinicians to upload CT, MRI, and X-ray scans, interact with the AI system through natural language queries, and receive diagnostic insights alongside gradient-based attribution maps rendered in an interactive visualization interface. RadEx AI demonstrates that the research contributions of this dissertation can be translated into a practically deployable clinical tool, bridging the gap between XAI research and clinical practice. The scientific

contributions of this work were further validated through a peer-reviewed publication accepted at an IEEE conference, confirming the novelty and rigor of the proposed approach within the international research community.

While this work firmly validates the proposed methodology, the experimental evaluation size is inherently bound to the available data for the specific model. Due to copyright restrictions, the evaluation was conducted on a focused set of volumetric images from the M3D-Cap and M3D-VQA datasets. By testing directly on this state-of-the-art model, the results are intrinsically confirmed and generalization is securely evaluated. Furthermore, while the teacher-forcing objective requires maintaining the full computation graph in memory during the backward pass, the subsequent computational overhead of Integrated Gradients is already well-known in the literature. It imposes heavy memory demands and multi-step path integration requirements, reflecting an intrinsic operational property of the established attribution method rather than a limitation of the proposed framework.

Several promising directions emerge from this work for future research. While generalization is established, the proposed attribution framework could undergo further evaluation on larger and more diverse datasets covering multiple imaging modalities and clinical specialties, in order to further establish the generalizability of the faithfulness hierarchy observed in this dissertation. A formal clinical user study involving practicing radiologists should be conducted to evaluate the usability and trust-building potential of the generated attribution maps in real diagnostic workflows. Furthermore, the integration of comprehensive clinical history and longitudinal patient data into the multimodal prompt architecture could significantly enhance the diagnostic aid provided to radiologists, mirroring the holistic evaluation process performed in clinical practice.

The results of this dissertation underscore a fundamental truth about the future of medical AI: accuracy alone is not sufficient for clinical adoption. As generative medical AI continues to advance, explainability will remain not an optional feature but the essential bridge between artificial intelligence and clinical trust. This work represents a step toward that bridge, demonstrating that even the most complex generative volumetric models can be made interpretable, and that interpretability and clinical utility can coexist in a single deployable system.

Scientific production



International Conference on
Electrical, Computer and Energy Technologies
(ICECET 2026)

March 10, 2026

ACCEPTANCE LETTER

Dear ***Abdelhalim, Youcef*; Ahmed Madani, Arif; Rachida, Saouli; Imane, Youkana***

Thank you for your submission to the International Conference on Electrical, Computer and Energy Technologies (ICECET 2026). We are pleased to inform you that your paper entitled "1459: **Post-Hoc Explainability for Generative Volumetric Medical Vision-Language Models**" has been accepted as a full paper for oral/poster/online presentation by the conference committee of ICECET 2026. The event will take place in Rome, Italy on 6-9 July 2026 online and physically.

We strictly follow "no podium, no paper" policy and only the papers that are presented at the conference will be submitted to IEEE Xplore for publication. **At least one author** of an accepted paper must register (as a full participant) and participate in ICECET 2026 online or physically for the paper to be included in the proceedings. You can make your registration and payment using this link: https://www.ecer.org/register/icecet

According to the conference regulations, only those papers which have been duly registered and presented on the conference days are considered for submission to IEEE Xplore. The conference program will be communicated in due course.

We look forward to seeing you for a fruitful research and innovation event and for a great time in the wonderful environment of Rome.

Yours sincerely,

Giancarlo Storti Gajani
Politecnico di Milano, Italy
ICECET 2026 Conference Chair

Web: https://www.icecet.com E-mail: info@icecet.com

Bibliography

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[3] F. Bai, Y. Du, T. Huang, and M.Q.-H. Meng. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

[4] M. Mersha et al. Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction. *Neurocomputing*, 2024.

[5] Geert Litjens et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[6] Humera Tariq and Syed Muhammad Burney. Brain mri literature review for interdisciplinary studies. *Journal of Biomedical Graphics and Computing*, 4(4):41–53, 10 2014.

[7] Kexin Wang et al. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2018.

[8] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[11] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2019.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[17] Runze Guo, Xiaojun Guo, Xiaoyong Sun, Peida Zhou, Bei Sun, and Shaojing Su. Background-aware cross-attention multiscale fusion for multispectral object detection. *Remote Sensing*, 16(21):4034, 10 2024.

[18] Alec Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[19] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *SSRN preprint*, 2024.

[20] Zifeng Wang, Zhenbang Wu, Dinesh Agrawal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and texts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

[21] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Besmira Nushi, Hoifung Poon, and Javier Alvarez-Valle. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022: 17th European Conference*, pages 1–21. Springer, 2022.

[22] Stefan Chen, Mandy Xu, Robert Yan, Stefan Harrer, and Curtis P. Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

[23] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung signatures Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 27854–27869, 2023.

[24] Y. Xin et al. Med3dvlm: An efficient vision-language model for 3d medical imaging. *arXiv preprint arXiv:2503.20047*, 2025.

[25] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[26] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[27] Ramprasaath Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[29] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[30] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado González, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, V. Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[31] Zhibo Zhang, Hussam Al Hamadi, Ernesto Damiani, Chan Yeob Yeun, and Fatma Taher. Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *arXiv preprint arXiv:2208.14937*, 8 2022.

[32] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations (ICLR)*, 2014.

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[34] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. In *Workshop at International Conference on Learning Representations (ICLR)*, 2015.

[35] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *PLOS ONE*, volume 10, 2015.

[36] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017.

[37] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey on explainable ai (xai): From principles to applications. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[38] David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. In *Advances in Neural Information Processing Systems*, 2018.

[39] Sangjoon Hwang et al. Visual language pre-training for medical image analysis. *arXiv preprint arXiv:2203.11111*, 2022.

[40] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Generating post-hoc explanation from deep neural networks for multi-modal medical image analysis tasks. *MethodsX*, 10:102009, 2023.

[41] Sara Gerke, Timo Minssen, and Glenn Cohen. Ethical and legal challenges of artificial intelligence-driven healthcare. *Artificial Intelligence in Healthcare*, pages 295–336, 2020.

[42] J. Hatherley et al. In defence of post-hoc explanations in medical ai. *Hastings Center Report*, 2025.

[43] Anees Ur Rahman Hashmi, Debesh Mahapatra, and Mohammad Yaqub. Envisioning MedCLIP: A deep dive into explainability for medical vision-language models. *arXiv preprint arXiv:2403.18996*, 2024.

[44] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

[45] G. Li, M. Zhang, Q. Zhang, and Z. Lin. Efficient binary 3D convolutional neural network and hardware accelerator. *Journal of Real-Time Image Processing*, 19(1):61–71, 2022.

[46] Xiaohua Zhai et al. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[47] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[48] Narine Kokhlikyan et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

[49] Robert W. Cox, John Ashburner, Helen Breman, Kate Fissell, Christian Haselgrove, Colin J. Holmes, Jack L. Lancaster, David E. Rex, Stephen M. Smith, Jeffrey B. Woodward, and Stephen C. Strother. A Sort of New Image Data Format Standard: NIfTI-1. *NeuroImage*, 22(1):e1440, 2004.

[50] Gordon Kindlmann. NRRD: Nearly Raw Raster Data. In *Proceedings of the IEEE Visualization Conference*, 1998.

[51] Luis Ibáñez, Will Schroeder, Lydia Ng, and Josh Cates. *The ITK Software Guide: The Insight Segmentation and Registration Toolkit*. Kitware, Inc., 2005.

[52] National Electrical Manufacturers Association. Digital imaging and communications in medicine (DICOM) standard. NEMA PS3, 2024.

الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي و البحث العلمي

جامعة محمد خيضر بسكرة

عنوان المشروع:

Explainable AI for Multi-Modality Deep Learning Models in Healthcare

مشروع لنيل شهادة مؤسسة ناشئة في اطار القرار الوزاري 1275

صورة العلامة التجارية



الاسم التجاري

RadEx

السنة الجامعية

2026_2025

بطاقة معلومات:

حول فريق الاشراف وفريق العمل

1- فريق الاشراف:

فريق الاشراف	
المشرف الرئيسي: رشيدة صولي (بن عبد العزيز)	التخصص: اعلام الي
المشرف المساعد (01): ايمان يوكانة	التخصص: اعلام الي
المشرف المساعد (2): العيد كحلول	التخصص: اعلام الي

2- فريق العمل:

فريق المشروع	التخصص	الكلية
الطالب: عبد الحليم يوسف	ذكاء اصطناعي	العلوم الدقيقة
الطالب: احمد مداني عريف	ذكاء اصطناعي	العلوم الدقيقة

عنوان المشروع : Explainable AI for Multi-Modality Deep Learning Models in Healthcare

فهرس المحتويات

المحور الأول: تقديم المشروع

المحور الثاني: الجوانب الابتكارية

المحور الثالث: التحليل الاستراتيجي للسوق

المحور الرابع: خطة الإنتاج والتنظيم

المحور الخامس :الخطة المالية

المحور السادس : النموذج الاولي التجريبي

المحور الأول: تقديم المشروع

1.1.فكرة المشروع (الحل المقترح)

-**مجال النشاط** :تنتمي RadEx إلى قطاع **الصحة الرقمية والذكاء الاصطناعي الطبي** تحديدًا تخصص الأشعة التشخيصية وهو قطاع خدمي تقني من الجيل الجديد يُعرَّف دوليًا بـ *E-health / HealthTech SaaS*.

-**كيف بدأت الفكرة وكيف تطورت؟** انطلقت الفكرة من تشخيص بنيوي دقيق لأزمة هيكلية في المنظومة الصحية الجزائرية :**شُح حاد في أطباء الأشعة** (0.8 طبيب لكل 100,000 نسمة مقابل 12 في فرنسا)، مما يُفضي إلى تأخر تشخيصي يتراوح بين 3 و8 أسابيع. ومن هذه الفجوة وُلدت الفرصة. تطورت الفكرة أكاديميًا إلى ورقة بحثية مقبولة في مؤتمر **ICECET 2026** بروما، ثم تجسّدت تقنيًا في نموذج أولي MVP مكتمل وقابل للعرض.

-**ما الذي ستقوم بهRadEx ؟** RadExمنصة ويب ذكية تُقدّم لأطباء الأشعة **مساعدًا تشخيصيًا قابلًا للتفسير** لا "صندوقًا أسود" عبر:

- تحليل صور الأشعة (CT, MRI) بـ **4 طرق تفسير متزامنة**
- توليد تقارير طبية آلية مع درجات ثقة
- الإجابة على الأسئلة التشخيصية (VQA)

-**كيف سيكون ذلك؟** من خلال نموذج **SaaS مؤسساتي (B2B)** بثلاث شرائح تسعير مصممة للسوق الجزائري (38 إلى 150 دولار/شهر)، مع بنية تقنية مبنية **بالكامل داخليًا** بصفر تبعيات خارجية ومستضافة على سحابة مؤهلة لحماية البيانات الصحية بمعيارHDS .

-**من الذي سيُنجز ذلك؟** فريق تأسيسي متكامل من مهندسَي ذكاء اصطناعي بإشراف أكاديمي متخصص:

- **عبد الحليم يوسف** CTO، طوّر محرك ال والمنصة بالكامل
- **عريف أحمد مداني** باحث أول، AI وتوثيق أكاديمي
- بإشراف **أ.د. رشيدة صولي ود. إيمان يوكانا, أ.د. العيد كحلول** من جامعة محمد خيضر بسكرة

-**أين سيتم إنجازه؟** يتمركز المشروع في **حاضنة المؤسسات الناشئة بجامعة محمد خيضر بسكرة**، ويستند علميًا وتقنيًا إلى **مختبر LINFI** (مختبر الإعلام والأنظمة الذكية) الذي تترأسه المشرفة الأكاديمية للمشروع أ.د. رشيدة صولي. هذه البيئة المؤسسية تمنح RadEx ثلاث مزايا استراتيجية متزامنة :**الغطاء الأكاديمي، والبنية التحتية البحثية، وشبكة الشراكات الجامعية** التي تُشكّل قناة تسويق مباشرة نحو المستشفيات الجامعية CHU وهي الشريحة الأولى المستهدفة في مرحلة الإطلاق.

1.2.القيم المقترحة

-**الحداثة** :RadEx تُعالج حاجة غير مُلبّاة بالكامل في السوق الجزائري والمغاربي. لا توجد حتى اليوم **أي منصة محلية أو إقليمية** متخصصة في الذكاء الاصطناعي القابل للتفسير (XAI) للأشعة الطبية. المنافسون الدوليون الوحيدون (Aidoc, Zebra, Qure.ai) غائبون كليًا عن السوق الجزائري، مما يجعل RadEx **عرضًا أصيلًا بلا سابقة محلية.**

-**الأداء** :تتفوق RadEx على المعيار الدولي أداءً وليس فقط سعرًا. بينما يعتمد المنافسون على **طريقة تفسيرواحدة**، تُشغّل RadEx **4 طرق تفسير متزامنة** مع آلية تحقق متقاطع.(Cross-Validation)

النتيجة الملموسة: سجَّل النموذج الأولي متوسط معالجة يبلغ 47 ثانية للصورة الواحدة، يُضاف إليها 2–4 دقائق للمراجعة الطبية، ليصبح الإجمالي 3–5 دقائق مقابل 15–20 دقيقة في القراءة اليدوية التقليدية (Baltruschat et al., Scientific Reports, 2019) ، أي رفع في الإنتاجية يتراوح بين 150 % و300 % حسب حجم العمل. هذا يتجاوز توقعات العميل، لا يساويها فحسب.

-**التكيّف** :صُمِّمت المنصة بـ 3**لوحات تحكم مستقلة** لكل شريحة عميل: لوحة القطاع الخاص (تركّز على السرعة) ، لوحة القطاع العمومي (مُحسَّنة لبيئة المستشفيات) ، ولوحة جامعية (للبحث والتعليم). هذا التخصيص البنيوي يعني أن المنتج **لا يتكيف مع العميل بل صُنع له منذ البداية.**

-**إنجاز المهمة** :مهمة طبيب الأشعة الجزائري اليوم هي تحليل 80 إلى 120 صورة يوميًا في ظل نقص حاد في الكوادر RadExتُنجز هذه المهمة بالتوازي مع تقديم: تحليل فوري، تقرير آلي جاهز، واقتراح إحالة متخصصة مما **يُحوّل الطبيب من مُنفّذ إلى مُراجِع.**

-**التصميم** :استُشير أطباء أشعة فعليون في تصميم المنصة قبل بناء أي سطر كود. لوحات التحكم الثلاث ليست خيارات جمالية بل استجابة مباشرة لبيئات عمل مختلفة جذريًا .

السعر: تتميز منصة RadEx بأسعار تنافسية تقل بشكل جذري عن أقرب الحلول الدولية المنافسة مثلQure.ai ، مع توفير وظائف متقدمة تشمل التفسير البصري(XAI) ، والإجابة عن الأسئلة الطبية(VQA) ، والتوليد الآلي للتقارير الطبية. تبدأ الاشتراكات الآن من 9,900 دج شهريًا (حوالي 42 دولارًا أمريكيًا وفق سعر الصرف الواقعي)، وهو استثمار يقل عن تكلفة فحصين بالأشعة في العديد من العيادات الخاصة، مما يتيح ولوجاً سهلاً للسوق ويضمن عائداً استثمارياً (ROI) فورياً للمؤسسات الطبية.
خفض التكاليف: تبلغ تكلفة معالجة صورة طبية واحدة عبر منصة RadEx حوالي 2.00 دج جزائري (0.008 دولار أمريكي تقريبًا) فقط، وذلك بفضل الاعتماد على بنية تحتية سحابية مستقلة وخوارزميات مطورة داخلياً تلغي الحاجة لتسديد رسوم الاستخدام للواجهات البرمجية (APIs) الجاهزة. تقارن هذه التكلفة الشحيحة بتكاليف تتراوح عادة بين 70 و700 دج للصورة الواحدة لدى الحلول الأجنبية المعتمدة على التسعير حسب حجم الاستخدام .(Per-Scan Pricing) هذا الهيكل المرن الذي يضمن للشركة الاحتفاظ بهامش ربح إجمالي ثابت يقدر بـ 80% يوفر للمؤسسات الصحية الجزائرية ترقية تكنولوجية ضخمة بأقل تكلفة تشغيلية ممكنة.

-**الحد من المخاطر** :يرسل المنافسون الدوليون بيانات المرضى إلى خوادم أجنبية (أمريكية، إسرائيلية، هندية) وهو ما يُعرّض المؤسسة الصحية لمخاطر قانونية جسيمة في ظل قانون حماية البيانات الجزائري .**18-07**

RadEx تُعالج الصور محليًا على استضافة سحابية مؤهلة صحيًا (HDS) **صفر تبعيات خارجية = صفر مخاطر امتثال.**

-**سهولة الوصول** :قبلRadEx ، لم يكن بإمكان مستشفى عمومي أو عيادة إقليمية في بسكرة أو ورقلة الوصول إلى أدوات ذكاء اصطناعي طبي بأي ثمن ببساطة لأنها غير متاحة في الجزائر RadEx تجعل هذه التقنية متاحة **لأي مؤسسة صحية جزائرية لديها اتصال بالإنترنت.**

-**سهولة الاستخدام** :المنصة مبنية كتطبيق ويب بواجهة مُبسَّطة لا تحتاج تثبيتًا، لا تدريبًا تقنيًا مطوّلًا، ولا بنية تحتية محلية. رفع الصورة، تحليل فوري، تقرير جاهز .**ثلاث خطوات.**

1.3.فريق العمل

- عبد الحليم يوسف: المؤسس المشارك والقائد التقني للمشروع، مهندس متخصص في الذكاء الاصطناعي والرؤية الحاسوبية، يشرف على تطوير نماذج التعلم العميق وتقنيات XAI الطبية وبنية منصة .RadEx

-عريف أحمد مداني: المؤسس المشارك والمسؤول البحثي، باحث في الذكاء الاصطناعي الطبي ونماذج اللغة الكبيرة، يساهم في الجانب الأكاديمي والتوثيق العلمي وتطوير النماذج البحثية .

-أ.د. رشيدة بن عبد العزيز : المشرفة الأكاديمية الأولى، أستاذة ورئيسة مخبر "الذكاء الاصطناعي والمعلوماتية" بجامعة محمد خيضر بسكرة، متخصصة في الذكاء الاصطناعي والتصوير الطبي، وتشرف على الجانب العلمي والبحثي للمشروع .

- د. إيمان يوكانا: مشرفة أكاديمية مساعدة، أستاذة محاضرة متخصصة في معالجة الصور والخوارزميات الموازية والذكاء الاصطناعي التطبيقي، تساهم في دعم الجانب التطبيقي وضمان توافق المنهجية البحثية مع السياق الطبي .

-د. لعيد كحلول: مشرف أكاديمي مساعد، يساهم في التأطير الأكاديمي والمتابعة العلمية للمشروع ودعم أنشطة البحث والتطوير المرتبطة بالمنصة.

1.4.أهداف المشروع

يهدف مشروع RadEx إلى تطوير منصة جزائرية متخصصة في الذكاء الاصطناعي القابل للتفسير في مجال الأشعة الطبية، من أجل تحسين سرعة ودقة التشخيص الطبي، وتقليل الضغط المتزايد على أطباء الأشعة داخل المؤسسات الصحية الجزائرية. ويرتكز المشروع على تقديم حل محلي منخفض التكلفة مقارنة بالمنصات الأجنبية، مع ضمان سيادة البيانات الصحية والامتثال لقانون حماية البيانات الجزائري.

الأهداف قصيرة المدى (0 إلى 12 شهر)

خلال السنة الأولى، تركز المؤسسة على استكمال مرحلة الإطلاق الفعلي للمنصة وتحقيق أولى النتائج التجارية والتقنية، وذلك من خلال:

- إطلاق النسخة التجريبية بالتعاون مع 3 مستشفيات جامعية شريكة .
- استقطاب ما بين 10 و20 مؤسسة صحية بين عيادات خاصة ومستشفيات ومراكز جامعية.
- تحقيق أولى الإيرادات المتكررة عبر نموذج الاشتراك SaaS
- الوصول إلى نقطة التعادل المالي قبل نهاية السنة الأولى وفق السيناريو الواقعي للدراسة المالية
- الحصول على Label Startup والانخراط ضمن منظومة الابتكار الصحي الرقمي في الجزائر .

الأهداف متوسطة المدى (12 إلى 24 شهر)

في المرحلة الثانية، تسعى المؤسسة إلى توسيع نشاطها التجاري والتقني داخل السوق الوطني، عبر:

- التعاقد مع ما بين 30 و50 مؤسسة صحية على المستوى الوطني.
- تحقيق إيرادات شهرية مستقرة تتجاوز 500,000 دج .
- توظيف فريق عمل متخصص يشمل التطوير، DevOps، والمبيعات .
- توسيع استخدام المنصة داخل القطاعين العمومي والجامعي .
- تطوير شراكات تقنية وأكاديمية جديدة تدعم النمو المستدام للمشروع .

الأهداف طويلة المدى (24 إلى 36 شهر)

على المدى البعيد، يهدف المشروع إلى التحول إلى فاعل إقليمي في مجال الذكاء الاصطناعي الطبي، من خلال:

- التوسع نحو أسواق المغرب العربي، خاصة تونس والمغرب .
- الحصول على شهادات واعتمادات دولية خاصة بالأجهزة الطبية الذكية .
- جذب جولة تمويل استثمارية لدعم التوسع التجاري والتقني .
- تطوير حلول إضافية تعتمد على الذكاء الاصطناعي الطبي متعدد الوسائط .
- ترسيخ مكانة RadEx كأول منصة مغاربية متخصصة في الأشعة الطبية القابلة للتفسير .

1.5.الحصة السوقية المستهدفة

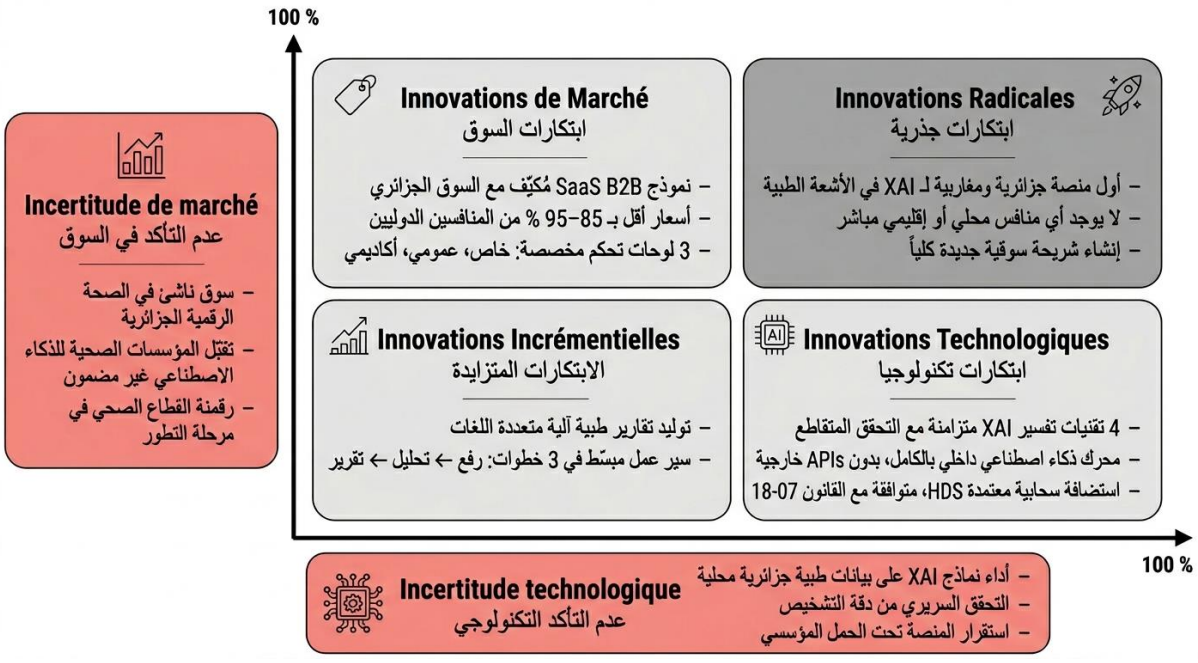
استناداً إلى إحصائيات وزارة الصحة والسكان وإصلاح المستشفيات (الكتاب السنوي للإحصاءات الصحية، 2025)، يُقدَّر السوق الإجمالي القابل للعنونة بحوالي 4300 مؤسسة ومختص في الأشعة داخل الجزائر، موزَّعين بين المستشفيات العمومية(EPH/EHS/CHU) ، والعيادات الخاصة المرخَّصة، وأطباء الأشعة المسجَّلين لدى الهيئات المهنية الوطنية. أما السوق القابل للخدمة، فيُقدَّر بحوالي 800 مؤسسة تمتلك الحد الأدنى من البنية التحتية الرقمية واتصالاً مستقراً بالإنترنت، وهو ما يمثّل نحو 18% من السوق الإجمالي، استناداً إلى تقارير رقمنة القطاع الصحي الصادرة عن وزارة الرقمنة والإحصائيات .(MPTIC, 2026) وفي هذا الإطار، تستهدف المؤسسة الوصول تدريجياً إلى ما بين 52 و90 مؤسسة صحية خلال السنوات الثلاث الأولى، أي ما يعادل حصة سوقية تتراوح بين 6.5% و11% من السوق القابل للخدمة، اعتماداً على وتيرة الرقمنة الصحية والشراكات الأكاديمية والطبية التي سيعتمد عليها المشروع.

1.6.جدول زمني لتحقيق المشروع



المحور الثاني: الجوانب الابتكارية

2.1.طبيعة الابتكارات



2.2.مجالات الابتكارات

يشمل مشروع **RadEx**عدة مجالات ابتكارية تجمع بين التكنولوجيا الحديثة والنموذج الاقتصادي المبتكر في قطاع الصحة الرقمية، ويمكن تلخيصها فيما يلي:

- **ابتكارات تكنولوجية :**

يعتمد المشروع على تطوير منصة ذكاء اصطناعي طبي قابلة للتفسير موجهة لتحليل صور الأشعة الطبية، باستخدام أربع تقنيات XAI متزامنة، مع توليد تقارير طبية ذكية متعددة اللغات. كما تم تطوير كامل البنية التقنية داخليًا دون الاعتماد على خدمات خارجية لمعالجة البيانات الطبية .

- **عمليات جديدة:**

يساهم المشروع في تحسين كفاءة العمل داخل أقسام الأشعة عبر تقليل مدة تحليل الصور الطبية وتسريع عملية إعداد التقارير التشخيصية، مما يساعد على رفع إنتاجية أطباء الأشعة وتقليل ضغط العمل داخل المؤسسات الصحية .

- **ميزات وخدمات جديدة:**

توفر المنصة خدمات غير متوفرة حاليًا في السوق الجزائري، مثل التفسير البصري لقرارات الذكاء الاصطناعي، الخرائط الحرارية الطبية، أنظمة الأسئلة والأجوبة الطبية الذكية(Medical VQA) ، ولوحات تحكم مخصصة للقطاع الخاص والعمومي والجامعي .

- **استهداف عملاء جدد:**

يستهدف المشروع فئات جديدة من المؤسسات الصحية التي لا تمتلك إمكانية الوصول إلى حلول الذكاء الاصطناعي الطبية الأجنبية بسبب ارتفاع تكلفتها أو مشاكل سيادة البيانات، مع التركيز على العيادات الخاصة، المستشفيات الجامعية، ومراكز البحث الطبي .

- **نماذج أعمال جديدة:**

يعتمد المشروع نموذج SaaS منخفض التكلفة ومتكيف مع السوق الجزائري، مع توفير اشتراكات مرنة وبرنامج "الشركاء التأسيسيين" لتسهيل تبني التكنولوجيا داخل المؤسسات الصحية .

- **ابتكارسوقي محلي:**

يمثل المشروع أول منصة جزائرية ومغاربية متخصصة في الذكاء الاصطناعي القابل للتفسير في مجال الأشعة الطبية، ما يمنحه أفضلية تنافسية داخل سوق لا يزال في بداياته محليًا وإقليميًا .

المحور الثالث: التحليل الاستراتيجي للسوق

3.1.عرض القطاع السوقي

السوق المحتمل

يتمثل السوق المحتمل لمشروع **RadEx** في جميع المؤسسات والأطراف التي يمكن أن تستفيد من حلول الذكاء الاصطناعي الموجهة للأشعة الطبية داخل الجزائر. ويشمل ذلك:

- المستشفيات العمومية .(EPH / EHS / CHU)
- العيادات الخاصة التي تحتوي على أقسام تصوير طبي .
- المراكز الجامعية وكليات الطب .
- أطباء الأشعة والمؤسسات البحثية الطبية .

وبحسب الدراسة السوقية الخاصة بالمشروع، يُقدّر السوق الإجمالي القابل للعنونة بحوالي **4300مؤسسة وطبيب أشعة محتمل** داخل الجزائر، مع توقع نمو هذا السوق بالتزامن مع تسارع رقمنة القطاع الصحي.

السوق المستهدف

يركز المشروع في مرحلته الأولى على المؤسسات الصحية التي تمتلك بنية رقمية مناسبة وقابلية لاعتماد الحلول الذكية، خاصة:

- العيادات الخاصة ذات النشاط المرتفع في الأشعة الطبية .
- المستشفيات الجامعية ومراكز البحث الطبي .
- المؤسسات الصحية التي تعاني من ضغط كبير في معالجة صور الأشعة .
- الجهات التي تبحث عن حلول منخفضة التكلفة مقارنة بالحلول الأجنبية .

ويُقدّر السوق القابل للخدمة بحوالي **800 مؤسسة صحية** تمتلك الحد الأدنى من الجاهزية الرقمية والاتصال المناسب بالإنترنت.

أسباب اختيارهذه الشريحة

تم اختيار هذه الفئة السوقية لعدة أسباب استراتيجية، أهمها:

- وجود نقص حاد في عدد أطباء الأشعة داخل الجزائر .
- الحاجة إلى تقليل مدة التشخيص وتسريع إعداد التقارير الطبية .
- ارتفاع تكلفة الحلول الأجنبية وصعوبة اعتمادها محليًا .
- تزايد الاهتمام برقمنة الخدمات الصحية داخل المؤسسات الجزائرية .
- حساسية البيانات الطبية وضرورة الحفاظ على سيادتها داخل الجزائر .

الحصة السوقية المستهدفة

يستهدف مشروع RadEx الوصول تدريجيًا إلى ما بين **52 و90 مؤسسة صحية** خلال أول ثلاث سنوات من النشاط، أي ما يعادل حصة سوقية تتراوح بين **6.5% و11%** من السوق القابل للخدمة، اعتمادًا على الشراكات الأكاديمية والطبية وسرعة تبني المؤسسات الصحية للحلول الرقمية الذكية.

3.2.قياس شدة المنافسة

يعمل مشروع **RadEx** داخل سوق يشهد وجود عدد محدود من المنافسين المحليين المباشرين في مجال تحليل الصور الطبية المعتمدة على الذكاء الاصطناعي القابل للتفسير، بينما تتمثل المنافسة الأساسية في المنصات الأجنبية المتخصصة في تحليل الصور الطبية.

المنافسون المباشرون

تشمل أبرز المنصات العالمية المنافسة:

- Aidoc
- Zebra Medical
- Qure.ai

وتقدم هذه الشركات حلول ذكاء اصطناعي لتحليل صور الأشعة الطبية، لكنها تعتمد غالبًا على خوادم أجنبية وتكاليف مرتفعة نسبيًا مقارنة بالسوق الجزائري.

نقاط قوة المنافسين

من أبرز نقاط القوة التي تتمتع بها الشركات الدولية:

- خبرة عالمية وسابقة في الأسواق الطبية الدولية .
- امتلاك شهادات واعتمادات صحية دولية .
- انتشار واسع داخل مؤسسات صحية كبرى .
- موارد مالية وتقنية ضخمة .

نقاط ضعف المنافسين

رغم قوتهم السوقية، توجد عدة نقاط ضعف تشكل فرصة لمشروع:RadEx

- أسعار مرتفعة جدًا مقارنة بقدرة المؤسسات الصحية الجزائرية .
- غياب سيادة البيانات بسبب الاعتماد على خوادم أجنبية .
- محدودية التخصيص للسوق الجزائري واحتياجاته المحلية .
- اعتماد أغلب الحلول على أنظمة "الصندوق الأسود" مع ضعف قابلية التفسير .
- تعقيد عمليات الدمج والدعم التقني بالنسبة للمؤسسات المحلية .

الميزة التنافسية لRadEx

يتميز مشروع RadEx بعدة عناصر تمنحه قدرة تنافسية داخل السوق الجزائري، أهمها:

- أول منصة جزائرية ومغاربية متخصصة في XAI الطبي .
- تكلفة أقل بنسبة كبيرة مقارنة بالمنافسين الدوليين .
- تطوير داخلي كامل للمحرك دون الاعتماد على APIs خارجية .
- دعم سيادة البيانات والامتثال للتشريعات الجزائرية .
- توفير أربع تقنيات تفسير متزامنة بدل تقنية واحدة فقط .

شدة المنافسة داخل السوق

تُعتبر شدة المنافسة المحلية حاليًا متوسطة إلى ضعيفة بسبب غياب حلول جزائرية متخصصة في المجال، ما يمنح المشروع فرصة للدخول المبكر وبناء قاعدة عملاء قبل توسع المنافسين الدوليين بشكل أكبر داخل السوق الجزائري.

الجدول (11) مقارن بين RadEx والحلول المنافسة

المعيار	RadEx	حلول أجنبية (Aidoc / Qure.ai)	حلول PACS التقليدية	أدوات ذكاء اصطناعي عامة
تفسير قابل للشرح (XAI)	نعم	محدود	لا	لا
استضافة محلية للبيانات الصحية	نعم	غالباً خارج الجزائر	حسب المؤسسة	لا
توافق مع سيادة البيانات الجزائرية	نعم	محدود	جزئي	لا
دعم اللغة العربية	نعم	محدود	لا	محدود
التكلفة السنوية	منخفضة	مرتفعة جداً	مرتفعة	متغيرة
تخصيص حسب المؤسسة	نعم	محدود	نعم	لا
التكامل مع PACS	نعم	نعم	نعم	محدود
واجهة موجهة للقطاع الجزائري	نعم	لا	محدود	لا
نموذج اشتراك SaaS مرن	نعم	نعم	لا	نعم
إمكانية العمل مع الجامعات والبحث العلمي	نعم	محدود	محدود	نعم
الدعم المحلي	نعم	لا	محدود	لا
سرعة النشر والتبني	مرتفعة	متوسطة	بطيئة	متوسطة

3.3.الاستراتيجية التسويقية

يعتمد مشروع **RadEx** على استراتيجية تسويقية موجهة للمؤسسات الصحية، ترتكز على إبراز القيمة التقنية والاقتصادية التي تقدمها المنصة مقارنة بالحلول الأجنبية مرتفعة التكلفة.

عناصر الاستراتيجية التسويقية

- الاعتماد على التسويق المباشر B2B نحو العيادات الخاصة والمستشفيات الجامعية .
- المشاركة في المؤتمرات والفعاليات الطبية للتعريف بالمنصة .
- بناء شراكات مع أطباء الأشعة والمؤسسات الأكاديمية .
- استغلال برنامج "الشركاء التأسيسيين" لجذب أول العملاء عبر عروض تفضيلية .
- استخدام LinkedIn والقنوات الرقمية المهنية للوصول إلى المؤسسات الصحية .

سياسة التسعير

يعتمد المشروع نموذج اشتراك SaaS منخفض التكلفة ومكيف مع السوق الجزائري، ما يسمح للمؤسسات الصحية بالاستفادة من خدمات الذكاء الاصطناعي الطبي بأسعار أقل بكثير من المنافسين الدوليين.

العلاقة مع العملاء

يركز المشروع على بناء علاقة مباشرة مع العملاء عبر:

- الدعم التقني المستمر .
- تخصيص بعض الخصائص حسب احتياجات المؤسسة الصحية .
- جمع الملاحظات لتحسين المنصة بشكل دوري .
- توفير تجربة استخدام بسيطة وسريعة داخل بيئة العمل الطبية .

تحليل SWOT



تحليل PESTEL

العوامل السياسية(Political)

- تولي الدولة الجزائرية اهتمامًا متزايدًا برقمنة قطاع الصحة ودعم الابتكار التكنولوجي، خاصة بعد جائحة COVID-19 ، ما يخلق بيئة داعمة للحلول الصحية الرقمية .
- توفر برامج حكومية داعمة للمؤسسات الناشئة مثل Startup Label و Algeria Venture و ANADEيمثل فرصة مهمة لتمويل وتوسيع المشروع .
- في المقابل، تبقى بعض الإجراءات الإدارية داخل القطاع الصحي العمومي بطيئة نسبيًا، ما قد يؤثر على سرعة التعاقد واعتماد الحلول الجديدة .

العوامل الاقتصادية(Economic)

- ارتفاع ميزانية قطاع الصحة في الجزائر يعزز فرص الاستثمار في الحلول الطبية الرقمية .
- السوق الجزائري ما يزال يعاني من نقص الحلول الطبية الذكية المحلية، ما يمنح RadEx فرصة للدخول بتكلفة أقل مقارنة بالمنافسين الدوليين .
- محدودية القدرة الشرائية لبعض المؤسسات الصحية تجعل نموذج التسعير المنخفض للمشروع عاملًا تنافسيًا مهمًا .
- الاقتصاد الرقمي الجزائري لا يزال في مرحلة النمو، لكنه يشهد تطورًا تدريجيًا في استخدام الخدمات الرقمية والإنترنت .

العوامل الاجتماعية(Social)

- يعاني القطاع الصحي الجزائري من ضغط متزايد ونقص في بعض التخصصات الطبية مثل الأشعة، ما يزيد الحاجة إلى أدوات مساعدة للتشخيص .
- هناك اهتمام متزايد داخل المجتمع الطبي بالحلول الرقمية التي تساعد على تحسين سرعة ودقة الخدمات الصحية .
- ما يزال جزء من المؤسسات الصحية والأطقم الطبية متحفظًا تجاه الاعتماد الكامل على الذكاء الاصطناعي، ما يتطلب بناء الثقة والتوعية التدريجية .

العوامل التكنولوجية(Technological)

- يشهد قطاع الصحة في الجزائر بداية تحول رقمي تدريجي يشمل السجلات الطبية الإلكترونية والأنظمة الرقمية للمستشفيات .
- لا تزال البنية التحتية الرقمية متفاوتة بين المؤسسات الصحية، خاصة خارج المدن الكبرى .
- يتميز مشروع RadEx بامتلاكه محرك ذكاء اصطناعي مطور داخليًا وعدم اعتماده على APIs خارجية، ما يعزز الاستقلالية والسيادة التقنية .
- تطور خدمات الحوسبة السحابية والـ GPU Cloud يساهم في تقليل تكاليف التشغيل مقارنة بالحلول التقليدية .

العوامل القانونية(Legal)

- يخضع المشروع لقانون حماية البيانات الشخصية الجزائري رقم 18-07، ما يفرض متطلبات صارمة فيما يتعلق بمعالجة البيانات الطبية الحساسة .
- تزايد الاهتمام بالأمن السيبراني في القطاع الصحي يفرض ضرورة تعزيز حماية البيانات والبنية التحتية الرقمية .
- الحصول على الاعتمادات والشهادات الطبية مستقبلاً سيكون عنصرًا أساسيًا لدعم التوسع المحلي والدولي للمشروع .

العوامل البيئية(Environmental)

- اعتماد المنصة على بنية رقمية سحابية يقلل الحاجة إلى البنية الورقية التقليدية داخل المؤسسات الصحية .
- الحلول الرقمية تساهم في تقليل التنقلات غير الضرورية وتسريع تبادل المعلومات الطبية بين المؤسسات .

المحور الرابع :خطة الإنتاج والتنظيم

4.1.عملية الانتاج

يمر تقديم خدمة RadEx بمراحل إنتاج رقمية تشمل استلام الصور الطبية وتحليلها وإصدار التقارير.



التموين

يعتمد مشروع **RadEx** على مزودين حقيقيين في مجال الحوسبة السحابية والبنية التحتية الرقمية، بما يتناسب مع طبيعة المنصة كحل ذكاء اصطناعي طبي سحابي.

سياسة الشراء

تشمل المشتريات والخدمات الأساسية:

- خوادم GPU لمعالجة وتحليل الصور الطبية .
- خدمات الاستضافة السحابية والتخزين .
- شراء نطاقات وأدوات الحماية والأمن السيبراني .
- تجهيزات تقنية أساسية للفريق (حواسيب وشاشات) .

الموردون المعتمدون

يتعامل المشروع مع مزودين معروفين في السوق التقني، من بينهم:

- **OVHcloud** لخدمات GPU Cloud والاستضافة السحابية .
- **Hetzner Cloud** كحل منخفض التكلفة للخوادم والبنية الخلفية .

- **Cloudflare** للحماية والأمان وإدارة النطاقات .
- مزودو تجهيزات محليون في الجزائر لاقتناء الحواسيب والشاشات .

سياسة الدفع والاستلام

يعتمد المشروع أساسًا على:

- اشتراكات شهرية أو سنوية للخدمات السحابية .
- الدفع الإلكتروني الدولي للخدمات التقنية الأجنبية .
- اقتناء التجهيزات المحلية حسب الحاجة التشغيلية .
- مرونة تغيير المزودين حسب الأداء والتكلفة ومتطلبات حماية البيانات الصحية .

4.2.اليد العاملة

يساهم مشروع **RadEx** في خلق مناصب عمل مباشرة وغير مباشرة داخل قطاع الصحة الرقمية والذكاء الاصطناعي في الجزائر.

مناصب العمل المباشرة

في المرحلة الأولى، يحتاج المشروع إلى فريق تقني مصغر يتكون من:

- مهندس ذكاء اصطناعي وتعلم عميق .
- مطور Backend و .DevOps
- مطور Frontend وتجربة مستخدم .
- مسؤول دعم تقني وعلاقات مع المؤسسات الصحية .

ومن المتوقع أن يوفر المشروع بين **4 و8 مناصب مباشرة** خلال أول سنتين حسب وتيرة التوسع.

مناصب العمل غير المباشرة

يساهم المشروع كذلك في خلق فرص عمل غير مباشرة في مجالات:

- الاستضافة والخدمات السحابية .

- التسويق الرقمي والمبيعات .
- التدريب التقني والصحي .

طبيعة اليد العاملة المطلوبة

يعتمد المشروع أساسًا على كفاءات مؤهلة في:

- الذكاء الاصطناعي وتحليل الصور الطبية .
- تطوير البرمجيات السحابية .
- الأمن السيبراني وحماية البيانات .
- إدارة الأنظمة والبنية التحتية الرقمية .

مصادر استقطاب الكفاءات

يستهدف المشروع استقطاب الكفاءات من:

- الجامعات الجزائرية، خاصة تخصصات الذكاء الاصطناعي والإعلام الآلي .
- مخابر البحث والمراكز الجامعية الطبية .
- حاملي شهادات الماستر والدكتوراه في المجالات التقنية والصحية الرقمية .
- المطورين المستقلين والمتخصصين في تقنيات Cloud و AIداخل الجزائر .

4.3.الشراكات الرئيسية

يعتمد مشروع **RadEx** على بناء شراكات استراتيجية مع مؤسسات تقنية، أكاديمية وصحية من أجل دعم تطوير المنصة وتسريع اعتمادها داخل السوق الجزائري.

الشركاء الأكاديميون والصحيون

- مستشفيات جامعية (CHU) من أجل التجارب الأولية والتحقق الطبي .
- كليات الطب ومخابر البحث الجامعية لدعم الجانب العلمي والبحثي .
- أطباء أشعة مختصون للمساهمة في اختبار وتحسين المنصة .

الشركاء التقنيون

- **OVHcloud** و **Hetzner Cloud** لتوفير خدمات الاستضافة والحوسبة السحابية .
- **Cloudflare** للحماية والأمن السيبراني وإدارة البنية الشبكية .
- مزودو خدمات GPU Cloud لمعالجة الصور الطبية بالذكاء الاصطناعي .

شركاء الدعم والتمويل

- **Algeria Venture** لدعم التمويل والمرافقة .
- برنامج **Startup Label Algeria** للاستفادة من الامتيازات الجبائية والتنظيمية .
- **ANADE** وبرامج دعم المؤسسات الناشئة في الجزائر .

أهمية هذه الشراكات

تساعد هذه الشراكات على:

- تسريع تطوير واعتماد المنصة داخل المؤسسات الصحية .
- ضمان بنية تقنية مستقرة وآمنة .
- تعزيز المصداقية العلمية والطبية للمشروع .
- تسهيل الوصول إلى التمويل وبرامج المرافقة والتوسع .

المحور الخامس: الخطة المالية PLAN FINANCIER

5.1.التكاليف والأعباء

تستند البنية المالية لمشروع **RadEx** إلى نموذج تشغيل منخفض التكلفة يعتمد خلال السنة الأولى على مبدأ **Sweat Equity**، حيث يعمل المؤسسان دون رواتب ثابتة مقابل حصص أسهم داخل المشروع، مما يسمح بالحفاظ على مستوى منخفض جداً من التكاليف الثابتة الشهرية مقارنة بالمشاريع التقليدية في مجال الصحة الرقمية والذكاء الاصطناعي الطبي.

كما أن جميع النماذج التقنية ومحرك الذكاء الاصطناعي التفسيري (XAI Engine) تم تطويرها داخلياً بالكامل من طرف الفريق المؤسس دون الاعتماد على واجهات خارجية مدفوعة(APIs) ، وهو ما يساهم في تقليل التكاليف التشغيلية طويلة المدى، ورفع هامش الربحية، وضمان سيادة البيانات الطبية الجزائرية داخل بنية سحابية مؤهلة وآمنة.

الجدول(01): البنية التحتية التقنية السنة الأولى (تقديرات مرجعية)

العنصر	المواصفات	التكلفة الشهرية (دج)	التكلفة السنوية (دج)	المرجع
GPUسحابي للاستدلال	NVIDIA L4أو ما يعادله، مزود مؤهل للبيانات الصحية	35,400دج (~$150)	424,800دج	المرجع: 236 دج/دولار (موازي 2026)
خادم التطبيق (Web App)	4 vCPU, 8GB RAM	5,900دج (~$25)	70,800دج	أسعار السوق، 2026
تخزين الصور الطبية	تخزين سحابي مؤهل(2TB)	3,540دج (~$15)	42,480دج	أسعار السوق، 2026
نطاقSSL +	radex.dz + Let's Encrypt	1,000دج	12,000دج	Dznic.dz، 2026
حماية CDN + (DDoS)	Cloudflare	مجاني	0	Cloudflare، 2026
قاعدة البيانات	PostgreSQLمدمجة داخل الخادم	مدمج	0	—
مخصص تقلبات سعر الصرف	احتياطي لمواجهة تغيّر سعر الدولار	7,000دج	84,000دج	تقدير داخلي
التحضير التنظيمي والامتثال	استشارات قانونية وتنظيمية أولية (Medical AI)	15,000دج	180,000دج	تقدير داخلي
الإجمالي الشهري	—	67,840دج	—	—

الجدول (02): التكاليف التشغيلية الشهرية السنة الأولى نموذج Sweat Equity

البند	التكلفة الشهرية (دج)	ملاحظات
البنية التحتية السحابية	45,840دج	GPUسحابي + خادم + تخزين + نطاق (راجع الجدول 01)
التسويق الرقميB2B	20,000دج	LinkedIn Ads + مؤتمرات طبية محلية
المصاريف الإدارية والقانونية	10,000دج	محاسب + تأمينات
التحضير التنظيمي والامتثال	15,000دج	استشارات قانونية وتنظيمية أولية (Medical AI)
احتياطي تقلبات سعر الصرف	7,000دج	احتياطي لمواجهة ارتفاع تكاليف الخدمات السحابية بالدولار الموازي
احتياطي المخاطر التشغيلية	9,784دج	10%من مجموع التكاليف التشغيلية قبل الاحتياطي (97,840 دج)
رواتب المؤسسين	0دج	— Sweat Equityمقابل حصص أسهم، لا رواتب ثابتة في السنة الأولى
الإجمالي الشهري	107,624دج(~456 USD)	—

يعمل مشروع **RadEx** خلال السنة الأولى **107,624 دج شهريًا** دون احتساب أي رواتب للمؤسسين(Sweat Equity) ، وهو مبلغ يعادل تقريبًا قيمة اشتراك مؤسسة صحية واحدة ضمن خطة **Professional** كما يتضمن هذا المبلغ مخصصات للامتثال التنظيمي واحتياطيًا لتقلبات سعر الصرف، مما يجعل التقديرات المالية أكثر تحفظًا وواقعية. ورغم ذلك، يظل النموذج الاقتصادي للمشروع قادرًا على بلوغ نقطة التعادل بعدد محدود من العملاء بفضل انخفاض التكاليف التشغيلية وارتفاع هامش الربح الإجمالي.

أولاً: التكاليف الثابتة (Charges Fixes)

تمثل التكاليف الثابتة النفقات التي تتحملها المؤسسة بغض النظر عن عدد العملاء أو حجم الصور الطبية المعالجة شهرياً. ويتميز نموذج **RadEx** بانخفاض هذه التكاليف بشكل استثنائي نتيجة الاعتماد على بنية تحتية سحابية مرنة، وعدم وجود رسوم متغيرة مرتبطة باستخدام خدمات ذكاء اصطناعي خارجي

الجدول (03): التكاليف الثابتة الشهرية السنة الأولى

البند	التكلفة الشهرية (دج)	التكلفة السنوية (دج)	الطبيعة
GPU سحابي للاستدلال	35,400 دج (~$150)	424,800 دج	ثابت
خادم التطبيق	5,900 دج (~$25)	70,800 دج	ثابت
تخزين سحابي (2TB)	3,540 دج (~$15)	42,480 دج	ثابت
نطاق SSL	1,000 دج	12,000 دج	ثابت
حماية DDoS	0	0	ثابت
قاعدة البيانات PostgreSQL	0	0	ثابت
التسويق الرقمي B2B	20,000 دج	240,000 دج	ثابت
المصاريف الإدارية والقانونية	10,000 دج	120,000 دج	ثابت
التحضير التنظيمي والامتثال	15,000 دج	180,000 دج	ثابت
احتياطي تقلبات سعر الصرف	7,000 دج	84,000 دج	ثابت
رواتب المؤسسين	0	0	Sweat Equity
إجمالي التكاليف الثابتة قبل احتياطي المخاطر	97,840 دج	1,174,080 دج	—
احتياطي المخاطر التشغيلية (10%)	9,784 دج	117,408 دج	ثابت
الإجمالي الكلي للتكاليف الثابتة	107,624دج(~456 USD)	1,291,488 دج(~5,472 USD)	—

الجدول (04): تطور التكاليف الثابتة خلال السنوات الثلاث الأولى

البند	السنة 1 (دج)	السنة 2 (دج)	السنة 3 (دج)	ملاحظات
البنية التحتية السحابية	550,080	800,000	1,200,000	تتوسع مع نمو قاعدة العملاء وارتفاع أحجام المعالجة السحابية
التسويق والمبيعات	240,000	480,000	720,000	إنشاء فريق مبيعات ابتداءً من السنة الثانية
المصاريف الإدارية والقانونية	120,000	240,000	360,000	محاسب + دعم قانوني
التحضير التنظيمي والامتثال	180,000	300,000	450,000	ملفات الامتثال التنظيمي، التحقق السريري، والتحضير للاعتمادات الطبية
إدارة مخاطر سعر الصرف	84,000	120,000	150,000	احتياطي لمواجهة تقلبات تكاليف الخدمات السحابية المسعّرة بالدولار
الرواتب	0	1,200,000	2,400,000	توظيف تدريجي بدءًا من السنة الثانية Sweat Equityفي السنة الأولى
الاستهلاكات(Amortissements)	100,000	120,000	150,000	إهلاك تجهيزات وأصول المنصة على أساس خطي (أصول إجمالية 870,000 دج)
إجمالي قبل احتياطي المخاطر	1,274,080	3,260,000	5,430,000	—
احتياطي المخاطر التشغيلية (10%)	117,408	314,000	528,000	10%من النفقات التشغيلية
الإجمالي السنوي الكلي	1,391,488	3,574,000	5,958,000	—

ثانياً: التكاليف المتغيرة (Charges Variables)

تظل التكاليف المتغيرة في مشروع **RadEx** منخفضة نسبيًا، ولا تتجاوز **20% من رقم الأعمال**، بفضل اعتماد المنصة على بنية تحتية شبه ثابتة وعدم استخدام خدمات ذكاء اصطناعي مدفوعة حسب عدد الطلبات. وتشمل هذه التكاليف أساسًا التخزين الإضافي، عرض النطاق الترددي، والتوسع التدريجي في الموارد السحابية مع نمو حجم الاستخدام.

الجدول(05): التكاليف المتغيرة لكل عميل

الطبيعة	الأثر الشهري التقديري	المكافئ بالدينار (236 دج/دولار)	التكلفة التقديرية	مصدر التكلفة المتغيرة
متغير روتيني	متوقع بانتظام عند تجاوز 7 عملاء نشطين يُغطّى ضمن التسعير الاشتراكي	3.5~دج/GB	0.015~دولار /GB إضافي	تخزين سحابي إضافي فوق TB2
متغير حسب الاستخدام	يرتفع مع كثافة رفع صور DICOMمُدرج ضمن هامش التكاليف المتغيرة	2.4~دج/GB	0.01~دولار/GB	عرض النطاق الترددي الزائد
متغير تدريجي	تنعكس في نمو بند البنية التحتية بالجدول 04	—	تزداد تدريجياً مع ارتفاع عدد الفحوصات والعملاء	توسعة الموارد السحابية عند زيادة الأحجام
متغير حسب المشروع	تُفعَّل عند التعاقد مع مستشفيات CHUأو مؤسسات تتطلب امتثالاً معززاً	—	تُسعَّر ضمن عقود المؤسسات الكبيرة	خدمات الامتثال والاستضافة الصحية المتقدمة (HDS)
متغير نادر	مُغطّى بالاحتياطي المخصص في الجداول 03 و04	—	نادر — مُدرج ضمن احتياطي المخاطر التشغيلية	معالجة استثنائية (أحجام دفعات غير اعتيادية)
—	836,640 دج (السنة 1) مُقيّدة استراتيجياً للحفاظ على هامش ربح 80%	—	20%من رقم الأعمال	نسبة التكاليف المتغيرة الإجمالية المعتمدة في الدراسة

ملاحظة جوهرية

يصل الهامش الإجمالي في مشروع **RadEx**إلى حوالي **80%**، وهو هامش مرتفع مقارنةً بمعظم حلول الذكاء الاصطناعي الطبية التجارية. ويرجع ذلك إلى أن الجزء الأكبر من البنية التحتية والتطوير يتم احتسابه ضمن التكاليف الثابتة، بينما تبقى الزيادة في التكاليف التشغيلية أقل من الزيادة في الإيرادات عند توسع قاعدة العملاء. ويُعد هذا النموذج الاقتصادي من أبرز نقاط القوة التنافسية للمشروع مقارنةً بالحلول المعتمدة على **APIs**مدفوعة حسب عدد الطلبات (**Per-Call Pricing**) والتي ترتفع تكلفتها بشكل مباشر مع نمو حجم الاستخدام.

الجدو ل(06): التكاليف المتغيرة حسب السيناريوهات

السيناريو	الإيرادات السنوية Y1 (دج)	التكاليف المتغيرة 20% (دج)	الهامش الإجمالي 80% (دج)
متحفظ (15 عميل ⟶ 8 نشط)	2,390,400	478,080	1,912,320
واقعي (24 عميل ⟶ 14 نشط)	4,183,200	836,640	3,346,560
متفائل (35 عميل ⟶ 22 نشط)	6,573,600	1,314,720	5,258,880

ثالثاً: تكلفة الوحدة معالجة صورة طبية واحدة

تُعد تكلفة معالجة الصورة الطبية الواحدة عبر منصة **RadEx** من أبرز المزايا التنافسية للمشروع، حيث تبلغ حوالي **13.4دج (0.056 USD~)**فقط للصورة الواحدة، مقارنة بتكاليف مرتفعة لدى المنصات الأجنبية المنافسة، والتي قد تصل إلى مئات الدنانير للصورة الواحدة. ويعود ذلك إلى تطوير محرك الذكاء الاصطناعي داخليًا واعتماد بنية تحتية سحابية منخفضة التكلفة مقارنة بالحلول التجارية المعتمدة على نماذج مدفوعة لكل طلب.

الجدول(07): تفصيل تكلفة معالجة صورة طبية واحدة

مرحلة المعالجة	الوصف	التكلفة المتغيرة والموزعة (دج)
التخزين السحابي ونقل البيانات	تخزين سحابي للصور الطبية وعرض النطاق	0.25~دج
تحليل XAI (تكلفة موزعة)	تشغيل محرك RadEx على GPU سحابي	1.50~دج
توليد التراكبات والتقارير	معالجة CPU عبر Sharp/OpenCV وتوليد النصوص	0.25~دج
احتياطي التوسع والامتثال	مخصص للتوسع السحابي والمتطلبات التنظيمية	(مدمج ضمن التقدير العام)
إجمالي التكلفة/صورة	—	2.00 ~دج(~0.008 USD)

الجدول(08): مقارنة تكلفة الوحدة مع المنافسين

المنصة	تكلفة/صورة(USD)	تكلفة/صورة (دج)	الفارق مقارنة بـRadEx
Aidoc	2.50 – 5.00 USD	590 – 1,180دج	أغلى بآلاف المرات
Zebra Medical	1.80 – 3.50 USD	425 – 825دج	أغلى بآلاف المرات
Qure.ai	0.50 – 1.20 USD	118 – 283دج	أغلى بآلاف المرات
RadEx	0.008 USD	2.00~دج	المرجع

الجدول(09): تكلفة الوحدة حسب حجم الاستخدام اليومي

ملاحظة الربحية	تكلفة التشغيل الموزعة (دج)	صور/شهر	حجم الاستخدام اليومي
هامش قوي للخطة Starter (9,900 دج	1,800~دج	900~صورة	عيادة صغيرة (30 فحص/يوم)
هامش مربح للخطة Professional (19,900دج	4,800~دج	2,400~صورة	عيادة متوسطة (80 فحص/يوم)
مغطاة باشتراك Institutional (39,900دج	12,000~دج	6,000~صورة	مستشفى عمومي (200 فحص/يوم)
تتوافق مع / Academic Research Lab	9,000~دج	4,500~صورة	مركز جامعي كبير (150 فحص/يوم)

رابعاً: خطط الاشتراك المفصَّلة

يعتمد مشروع **RadEx**على نموذج اشتراك **SaaS** متعدد الشرائح، يسمح بتكييف الخدمات حسب حجم المؤسسة الصحية وطبيعة الاستخدام. وتشمل الخطط خدمات تحليل الصور الطبية، التفسير البصري **XAI**، التوليد الآلي للتقارير الطبية، ولوحات التحكم المؤسسية، مع الحفاظ على أسعار تنافسية مقارنة بالحلول الأجنبية، مع مراعاة ارتفاع تكاليف البنية التحتية السحابية بعد اعتماد سعر السوق الموازية للدولار.

الجدول(10): الشريحة 1 العيادات الخاصة (Cliniques Privées)

الملف النمطي: عيادة أشعة خاصة، من 1 إلى 5 أطباء أشعة، وحجم نشاط يتراوح بين 30 و80 فحصاً يومياً.

الخطة	حجم التحليل	طرقXAI	VQA	التقارير	الدعم	السعر الشهري (دج)	السعر السنوي (دج)
Starter	1,000 تحليل/شهر	طريقتان أساسيتان	لا	تقرير أساسي	بريد إلكتروني	9,900	118,800
Professional	3,000 تحليل/شهر	4طرق كاملة	نعم	تقرير كامل + إحالة	أولوية	19,900	238,800
Premium	غير محدود	4طرق + تقارير مخصصة	نعم	مخصص + APIخاص	أولوية قصوى	29,900	358,800

الجدول(11): الشريحة 2 المستشفيات العمومية(EPH / CHU)

الملف النمطي: مستشفى عمومي أو مركز استشفائي جامعي، يضم من 5 إلى 15 طبيب أشعة، مع حجم نشاط يتراوح بين 100 و300 فحص يومياً، ويحتاج إلى حلول مركزية قابلة للتوسع والتكامل مع أنظمة.PACS

السعر السنوي (دج)	السعر الشهري (دج)	تدريب الفريق	تكامل PACS	لوحة التحكم	طرقXAI	حجم التحليل	الخطة
478,800	39,900	لا	لا	لوحة عمومية	4 طرق كاملة	5,000 تحليل/شهر	Institutional
718,800	59,900	نعم	نعم	لوحة مخصصة	4 طرق + إحصائيات متقدمة	غير محدود	Institutional Plus

الجدول(12): الشريحة 3 المراكز الجامعية والبحثية (CHU + Facultés de Médecine)

الملف النمطي: مركز جامعي طبي أو مخبر بحثي، يضم أكثر من 10 أطباء مقيمين وباحثين، مع استخدام بحثي وتعليمي مكثف لمنصة الذكاء الاصطناعي الطبي.

السعر السنوي (دج)	السعر الشهري (دج)	أدوات تعليمية	تصدير البيانات	لوحة التحكم	طرقXAI	حجم التحليل	الخطة
178,800	14,900	نعم	لا	لوحة جامعية	4 طرق كاملة	2,000 تحليل/شهر	Academic
358,800	29,900	نعم	نعم	مخصصة بحثياً	4 طرق API + بحثي	غير محدود	Research Lab

برنامج الشركاء التأسيسيين(Founding Partners Program)

لتسريع اكتساب العملاء وتأمين إيرادات مبكرة، يقدّم مشروع **RadEx** برنامجًا خاصًا موجّهًا لأول 10 عملاء مؤسساتيين، بهدف بناء مرجعية محلية قوية والحصول على بيانات تحقق سريري حقيقية داخل السوق الجزائري، مع تشجيع المؤسسات الصحية على تبني المنصة في مراحلها الأولى مقابل مزايا تفضيلية طويلة المدى.

الجدول(13): برنامج الشركاء التأسيسيين

التفاصيل	لميزة
15%على الاشتراك السنوي مقابل التزام لمدة 12 شهرًا	الخصم التأسيسي
أولوية في تطوير المميزات بناءً على احتياجات المؤسسة	أولوية التطوير
شهادة "شريك تأسيسي" لاستخدامها في المواد التسويقية	شهادة الشراكة
إيرادات مبكرة مضمونة + بيانات تحقق سريري حقيقية	الفائدة لـRadEx

الجدول(14): ملخص مقارن لجميع خطط الاشتراك

الفارق	المنافس الأقرب	السعر/سنة (دج)	السعر/شهر (USD)	السعر/شهر (دج)	الشريحة	الخطة
أرخص بكثير	Qure.ai (~500 USD/شهر	118,800	~42	9,900	عيادات خاصة	Starter
أرخص بكثير	Qure.ai (~1,000 USD/شهر	238,800	~84	19,900	عيادات خاصة	Professional
أرخص بكثير	Qure.ai (~1,500 USD/شهر	358,800	~126	29,900	عيادات خاصة	Premium
أرخص بكثير	Zebra Medical (~2,000 USD/شهر	478,800	~169	39,900	مستشفيات عمومية	Institutional
أرخص بكثير	Aidoc (~3,000 USD/شهر	718,800	~253	59,900	مستشفيات عمومية	Institutional Plus
—	لا يوجد منافس محلي مباشر	178,800	~63	14,900	مراكز جامعية	Academic
—	لا يوجد منافس محلي مباشر	358,800	~126	29,900	مراكز جامعية	Research Lab

5.2.رقم الأعمال (Chiffre d’Affaires)

يمثل رقم الأعمال مجموع الإيرادات الناتجة عن اشتراكات وخدمات منصة RadEx الموجهة للمؤسسات الصحية، ويعكس قدرة المشروع على تحقيق مداخيل متكررة وقابلة للنمو اعتمادًا على نموذج SaaS. وقد تم إعداد التوقعات المالية بالاعتماد على تطور عدد العملاء النشطين، ومتوسط الاشتراك السنوي، إضافة إلى الخدمات الإضافية المرتبطة بالتكامل مع أنظمة PACS، التكوين، والدعم التقني.

كما تم احتساب رقم الأعمال المتوقع بالاعتماد على عدد العملاء النشطين ومتوسط الإيراد الشهري لكل مؤسسة صحية، وذلك من خلال ضرب عدد المؤسسات المشتركة في متوسط الاشتراك الشهري ثم احتساب القيمة السنوية على مدار 12 شهرًا.

وبعد مراجعة خطط التسعير الجديدة وهيكلة مزيج العملاء المستهدف، تم اعتماد متوسط إيراد شهري يُقدَّر بحوالي 24,900 دج لكل مؤسسة صحية، أي ما يعادل تقريبًا 298,800 دج سنويًا لكل عميل نشط.

وترتكز هذه التوقعات على تطور عدد المؤسسات الصحية المشتركة تدريجيًا خلال السنوات الأولى، مع الاعتماد على نموذج اشتراك SaaS موجه للعيادات الخاصة، المستشفيات العمومية، والمراكز الجامعية والبحثية.

الجدول (15): رقم الأعمال المتوقع حسب السيناريوهات

السيناريو	عدد العملاء النشطين	متوسط الإيراد الشهري لكل عميل	رقم الأعمال السنوي (دج)
متحفظ	5عملاء	24,900دج	1,269,900
واقعي	8عملاء	24,900دج	1,980,840
متفائل	14عميل	24,900دج	4,183,200

الجدول (16): تطور رقم الأعمال خلال السنوات الثلاث الأولى

لبند	السنة 1	السنة 2	السنة 3
عدد العملاء النشطين	8	22	45
متوسط الإيراد الشهري لكل عميل	20,633دج	25,581دج	25,455دج
رقم الأعمال الشهري	165,070دج	562,800دج	1,145,500دج
رقم الأعمال السنوي	1,980,840دج	6,753,600دج	13,746,000دج
معدل النمو السنوي	—	%241	%103

DETAIL CHIFFRE D'AFFAIRES RadEx

المنتج / الخدمة الموجهة للعميل	N	N+1	N+2	N+3	N+4	N+5
خطةProfessional						
كمية الاشتراكات	5	12	25	45	65	90
السعر السنوي بدون ضرائب (دج)	238,800	238,800	238,800	238,800	238,800	238,800
المبيعات السنوية (دج)	1,014,900	2,865,600	5,970,000	10,746,000	15,522,000	21,492,000
خطةInstitutional						
كمية الاشتراكات	2	7	14	22	33	45
السعر السنوي بدون ضرائب (دج)	478,800	478,800	478,800	478,800	478,800	478,800
المبيعات السنوية (دج)	813,960	3,351,600	6,703,200	10,533,600	15,800,400	21,546,000
خطةAcademic						
كمية الاشتراكات	1	3	6	8	12	15
السعر السنوي بدون ضرائب (دج)	178,800	178,800	178,800	178,800	178,800	178,800
المبيعات السنوية (دج)	151,980	536,400	1,072,800	1,580,400	2,313,600	2,982,000
رقم الأعمال الإجمالي	1,980,840	6,753,600	13,746,000	22,860,000	33,636,000	46,020,000

5.3. جدول حسابات النتائج المتوقعة

يُمثّل هذا الجدول ملخصاً مالياً يوضح تطور إيرادات ومصاريف مشروع **RadEx** خلال السنوات الأولى من النشاط، وينتهي بتحقيق نتيجة مالية إيجابية (ربح) أو سلبية (خسارة) خلال فترة زمنية تُعرف بالسنة المحاسبية.

الجدول (17): حسابات النتائج المتوقعة

البند	السنة 1	السنة 2	السنة 3
الإيرادات — اشتراكات SaaS (دج)	1,980,840	6,753,600	13,746,000
تكلفة الخدمات المباعة(20%) COGS	(396,168)	(1,350,720)	(2,749,200)
الهامش الإجمالي (80%)	1,584,672	5,402,880	10,996,800
البنية التحتية السحابية	(550,080)	(800,000)	(1,200,000)
التسويق والمبيعات	(240,000)	(480,000)	(720,000)
المصاريف الإدارية	(120,000)	(240,000)	(360,000)
التحضير التنظيمي والامتثال	(180,000)	(300,000)	(450,000)
إدارة مخاطر سعر الصرف	(84,000)	(120,000)	(150,000)
مخصصات المخاطر التشغيلية	(117,408)	(314,000)	(528,000)
الرواتب Sweat Equity) في السنة الأولى(	0	(1,200,000)	(2,400,000)
الاستهلاكات المحاسبية	(100,000)	(120,000)	(150,000)
النتيجة التشغيليةEBIT	193,184	1,828,880	5,038,800

الجدول (18): احتياجات رأس المال العامل (BFR)

نصر الحساب	التفاصيل	القيمة (دج)
الإيراد الشهري المرجعي(CA mensuel)	8عملاء (مع خصم 15%)	165,070
حسابات القبض القطاع العمومي	165,070 × 30% × 2شهرين	99,042
حسابات القبض القطاع الخاص	165,070 × 70% × 1شهر واحد	115,549
إجمالي حسابات القبض	—	214,591
ديون الموردين (السحابة والخدمات)	مدفوعات شهرية مسبقة	(45,840)
إجماليBFR	حسابات القبض – ديون الموردين	407,340

الجدول (19): تفسير نتيجة BFR

لعنصر	التفاصيل
طبيعةBFR	موجب ومحدود، مما يعني أن المشروع يحتاج تمويلًا قصير المدى لتغطية فارق التحصيل
مصدر التغطية	الاحتياطي المدرج ضمن التمويل الأولي، إضافة إلى التدفقات النقدية الناتجة عن الاشتراكات المسبقة
عامل التخفيف	70%من العملاء من القطاع الخاص يدفعون مسبقًا أو خلال دورة تحصيل قصيرة، مما يقلص دورة التحصيل
مخاطر القطاع العمومي	إمكانية تأخر التحصيل إلى غاية 60 يومًا في المستشفيات العمومية / EPH CHU

تحليل نقطة التعادل(Seuil de Rentabilité)

الجدول (20): نقطة التعادل حسب عدد العملاء

القيمة	**نصر الحساب**
115,957دج	التكاليف الثابتة الشهرية (CF)
24,900دج	متوسط الإيراد لكل عميل (ARPU)
%80	نسبة الهامش الإجمالي
19,920دج	هامش المساهمة لكل عميل
6عملاء	نقطة التعادل

الجدول (21): نقطة التعادل حسب الإيرادات

القيمة	**نصر الحساب**
115,957دج	التكاليف الثابتة الشهرية
%80	نسبة الهامش الإجمالي
144,946دج / شهر	نقطة التعادل بالإيرادات الشهرية

الجدول (22): مزيج العملاء المعتمد في حسابARPU

نوع العميل	نسبة المزيج	متوسط الإيراد الشهري	الإيراد المرجح
عيادة خاصة(Professional)	%57.14	19,900دج	11,371دج
مستشفى عمومي(Institutional)	%28.57	39,900دج	11,399دج
مركز جامعي(Academic)	%14.29	14,900دج	2,130دج
متوسط الإيراد لكل عميل شهريًا (ARPU)	**%100**	**—**	**24,900دج**

الجدول (23): توقيت بلوغ نقطة التعادل حسب السيناريوهات

السيناريو	عدد العملاء نهاية السنة الأولى	متوسط العملاء النشطين	الإيرادات السنوية (دج)	النتيجة الصافية (دج)	توقيت بلوغ نقطة التعادل
متحفظ	8عملاء	5عملاء	1,245,000	-234,000	السنة الثانية
واقعي	12عميل	8عملاء	1,980,840	+193,184	الشهر 10 – 11
متفائل	24عميل	14عميل	4,183,200	+1,955,072	الشهر 5 – 6

الجدول (24): عوامل انخفاض نقطة التعادل

العامل	الأثر
نموذجSweat Equity	غياب الرواتب الثابتة للمؤسسين خلال السنة الأولى، مما يخفض النفقات التشغيلية بشكل ملحوظ.
محرك RadEx المطور داخليًا	عدم وجود رسوم استخدام لنماذج ذكاء اصطناعي خارجية مدفوعة حسب الطلب.
هامش إجمالي مرتفع(80%)	يسمح بتحويل الجزء الأكبر من الإيرادات إلى تدفقات نقدية بعد تغطية التكاليف المتغيرة.
إعفاء ضريبي Startup Label	إعفاء من الضريبة على الأرباح خلال السنوات الأولى وفق الامتيازات التنظيمية للمؤسسات الناشئة.
نقطة تعادل منخفضة	يحتاج المشروع إلى حوالي 7 عملاء نشطين فقط لتحقيق التوازن المالي.

5.3.خطة الخزينة

تُعدّ خطة الخزينة وثيقة مالية تسمح بتحديد جميع الإيرادات والنفقات المتوقعة شهرياً خلال السنة الأولى من نشاط مؤسسة **RadEx**، كما تساعد على متابعة تطور السيولة المالية وضمان قدرة المشروع على تغطية التزاماته التشغيلية بشكل مستمر.

البند	الشهر 1	الشهر 2	الشهر 3	الشهر 4	الشهر 5	الشهر 6	الشهر 7	الشهر 8	الشهر 9	الشهر 10	الشهر 11	الشهر 12	المجموع السنوي
المؤشرات التجارية													
عدد العملاء النشطين	0	1	2	3	5	7	9	11	12	13	14	14	—
التدفقات الداخلة													
تحصيلات العملاء (دج)	0	50,000	70,000	80,000	100,000	120,000	150,000	180,000	220,000	240,000	260,000	276,249	1,766,249
تمويل أولي ودعم استثنائي	1,200,000	0	0	0	0	0	0	0	0	0	0	0	1,200,000
إجمالي الإيرادات النقدية (دج)	**1,200,000**	**50,000**	**70,000**	**80,000**	**100,000**	**120,000**	**150,000**	**180,000**	**220,000**	**240,000**	**260,000**	**276,249**	**2,966,249**
التدفقات الخارجة													
البنية التحتية السحابية (دج)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(45,840)	(550,080)
تسديد التكاليف المتغيرة COGS (دج)	0	(10,000)	(15,000)	(16,000)	(20,000)	(24,000)	(30,000)	(36,000)	(44,000)	(48,000)	(52,000)	(53,168)	(348,168)
التسويق الرقمي B2B (دج)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(20,000)	(240,000)
المصاريف الإدارية والقانونية (دج)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(10,000)	(120,000)

الجدول (26): ملخص الخزينة السنوي

البند	القيمة السنوية (دج)
إجمالي تحصيلات العملاء	1,766,249
إجمالي التمويل والدعم التأسيسي	1,200,000
إجمالي الإيرادات النقدية السنوية	**2,966,249**
إجمالي النفقات والاستثمارات (متضمنة 870 ألف دج للتجهيزات)	(2,509,656)
مقاصة ديون الموردين والمستحقات المتبقية بنهاية السنة	(20,000)
صافي الخزينة في نهاية السنة الأولى	**+436,593**

المحور السادس : النموذج الاولي التجريبي

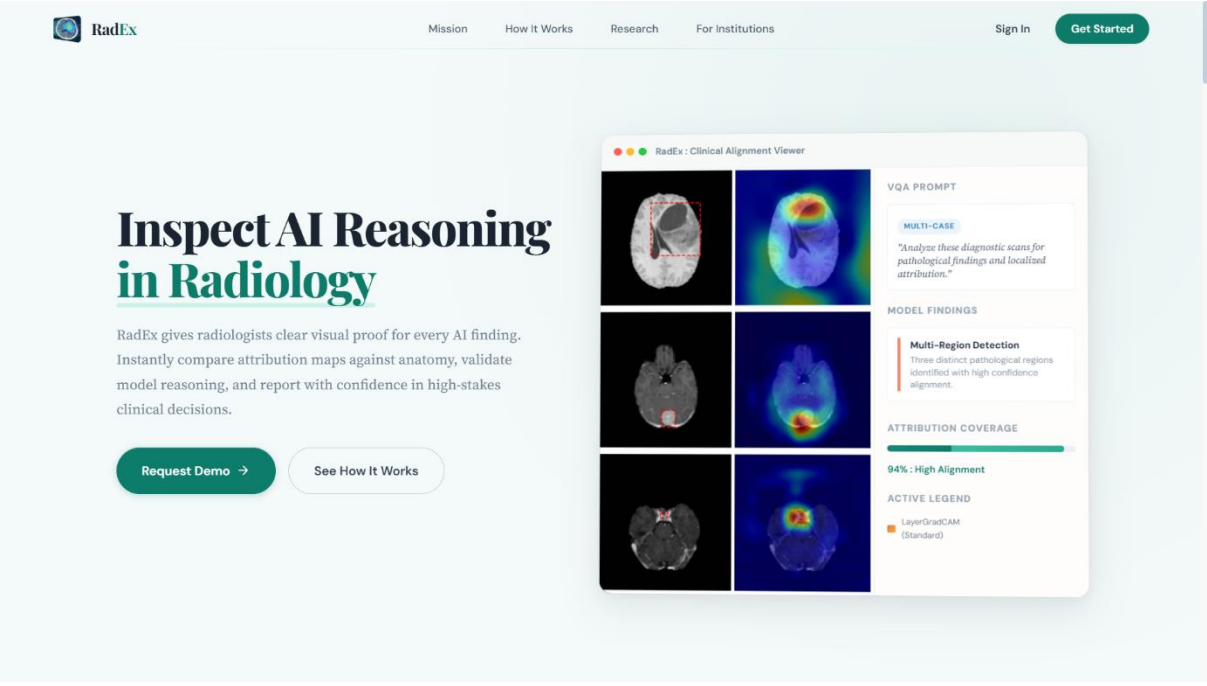
قام مشروع RadEx بتطوير نموذج أولي رقمي (MVP) لمنصة ذكاء اصطناعي متخصصة في تحليل الصور الطبية القابلة للتفسير، بهدف توفير حل عملي يساعد المؤسسات الصحية على تسريع التشخيص وتحسين دقة تحليل صور الأشعة الطبية. وقد تم تصميم المنصة لتلائم احتياجات مختلف الفئات المستهدفة، سواء المؤسسات الصحية الخاصة والعمومية أو الجهات الأكاديمية والتعليمية.

يعتمد النموذج الأولي على واجهات استخدام تفاعلية وبنية سحابية مرنة تسمح بإدارة الصور الطبية، تحليلها باستخدام الذكاء الاصطناعي، وتوليد تقارير تفسيرية بطريقة سهلة وآمنة. كما تم تطوير خصائص مختلفة حسب نوع المستخدم من أجل توفير تجربة استعمال مناسبة لكل فئة.

وفيما يلي عرض لأهم واجهات ووظائف المنصة من خلال لقطات متسلسلة توضح طريقة عمل النظام بالنسبة لمختلف أنواع المستخدمين.

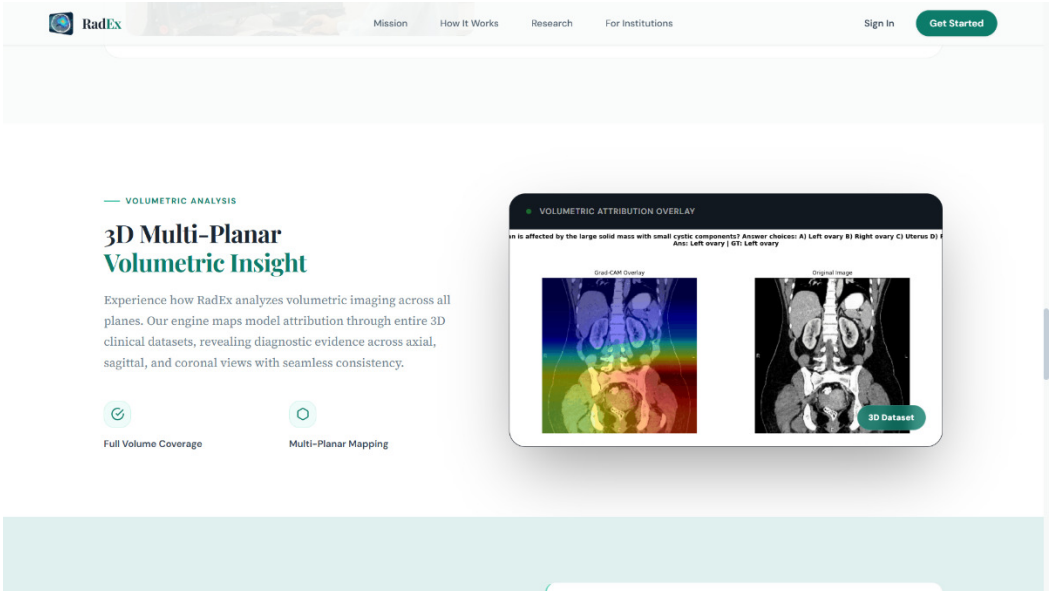
6.1.الصفحة الرئيسية للمنصة

تعرض هذه الواجهة نظرة عامة حول قدرات منصة RadEx في تحليل الصور الطبية القابلة للتفسير، مع تقديم أهم الخدمات والمزايا الموجهة للمؤسسات الصحية والبحثية.



6.2.التفسير البصري لقرارات الذكاء الاصطناعي

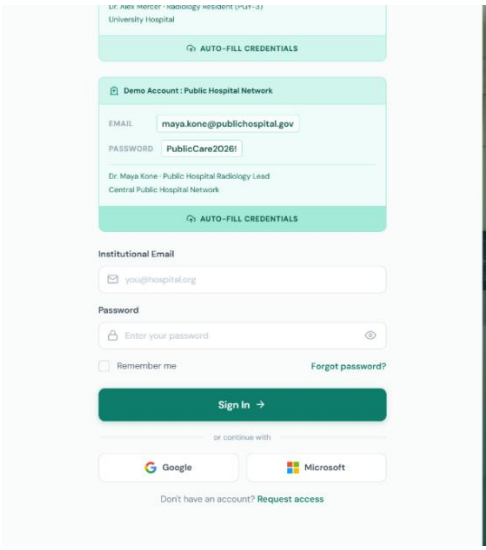
توضح هذه الواجهة كيفية عرض الخرائط التفسيرية (XAI Heatmaps) التي تساعد الأطباء على فهم أسباب قرارات الذكاء الاصطناعي وتحسين موثوقية التشخيص.

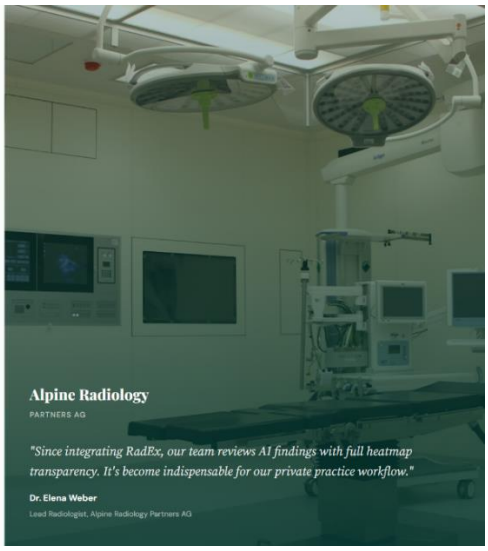


6.3.واجهات المؤسسات الصحية الخاصة والعمومية

1-واجهة تسجيل الدخول وإدارة الحسابات

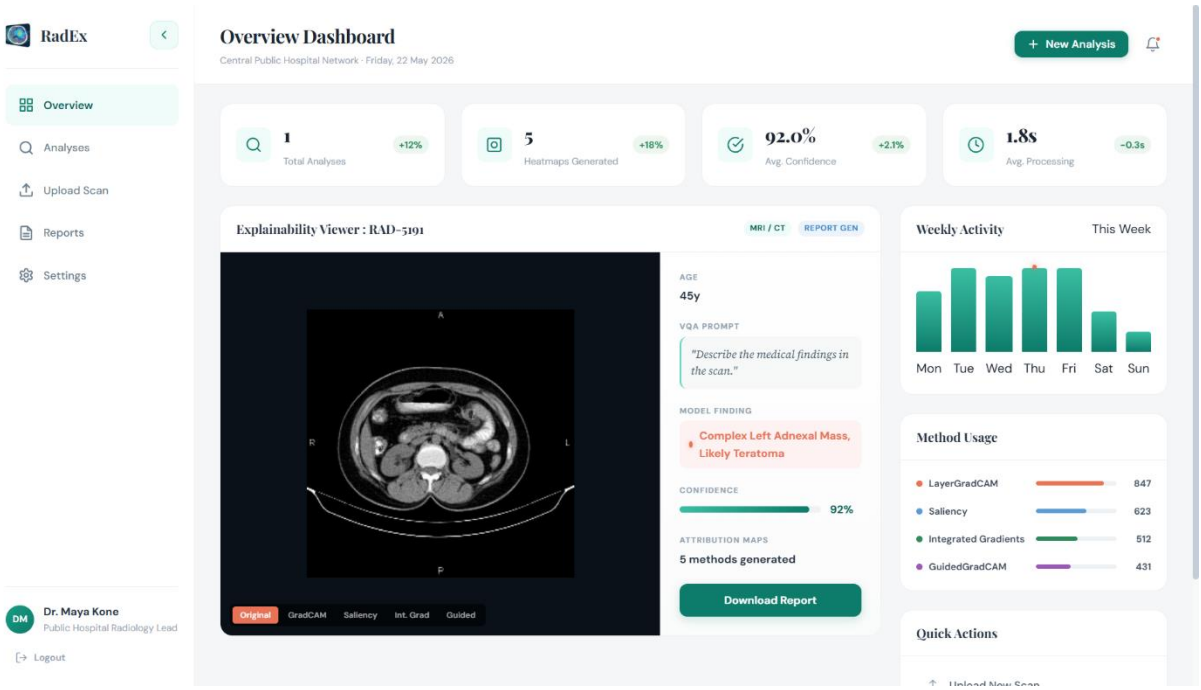
توضح آلية الدخول إلى المنصة باستخدام الحسابات المؤسسية وإدارة صلاحيات المستخدمين داخل المؤسسة الصحية.





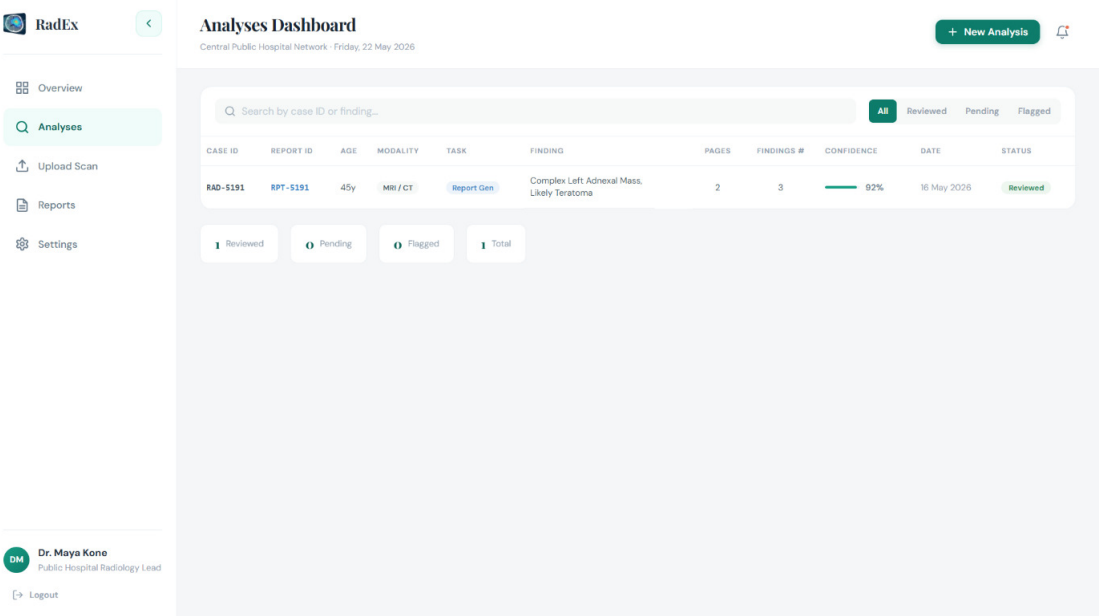
2-لوحة التحكم الرئيسية والإحصائيات

تعرض نظرة عامة حول النشاطات الطبية، عدد التحليلات المنجزة، نسب الثقة، وطرق التفسير المستخدمة.



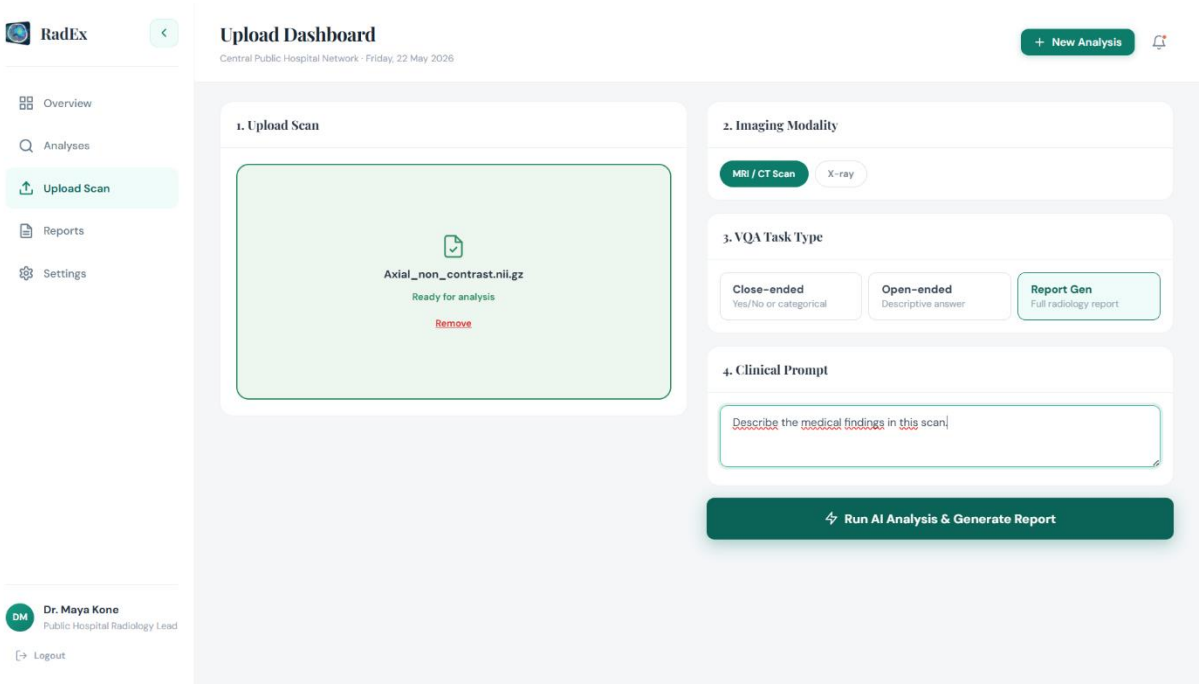
3-واجهة إدارة التحاليل الطبية

تمكّن المستخدم من استعراض جميع التحاليل المنجزة، حالة التقارير، ومستوى الثقة الخاص بكل تحليل.



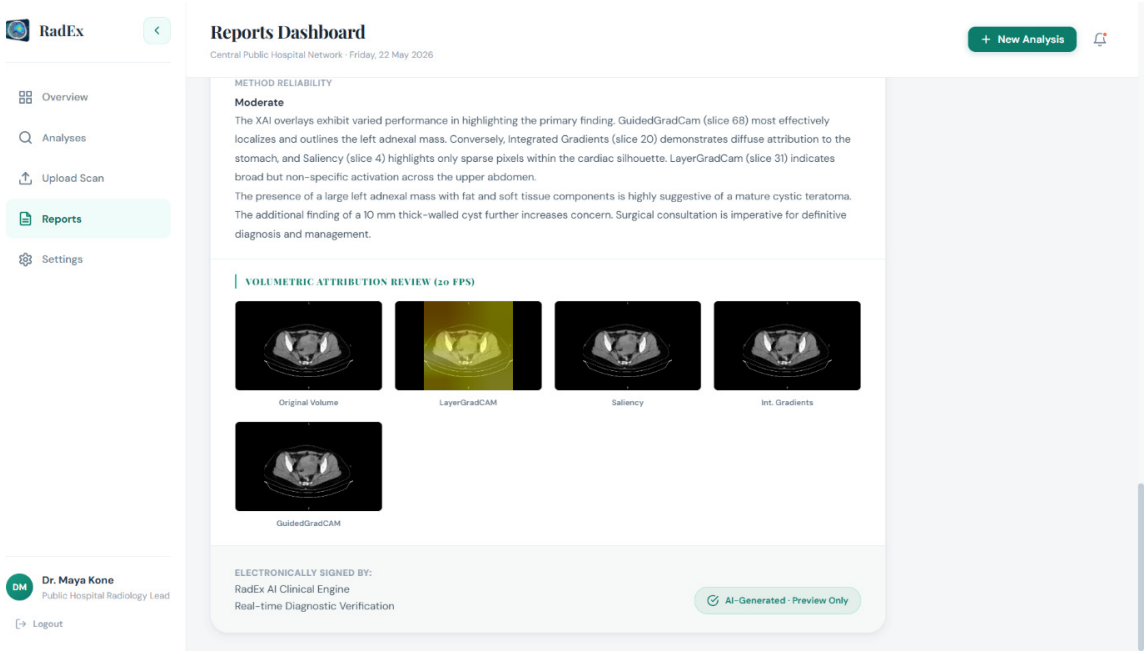
4-واجهة رفع صور الأشعة الطبية

تسمح برفع ملفات التصوير الطبي واختيار نوع الفحص وطبيعة المهمة الطبية المطلوبة.



5-توليد التقارير الطبية الذكية

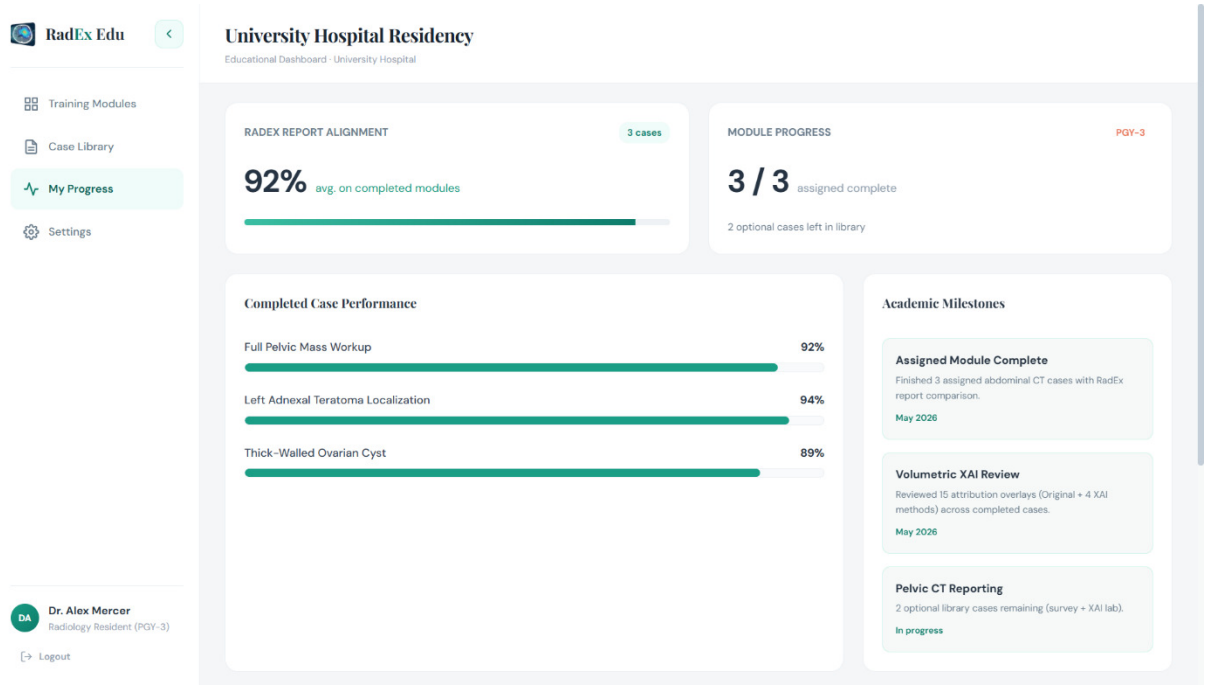
يتم إنشاء تقرير طبي تفصيلي بشكل آلي يتضمن النتائج السريرية، مؤشرات الثقة، والتفسير الطبي للحالة.



6.4.واجهات المستخدمين الأكاديميين والتعليميين

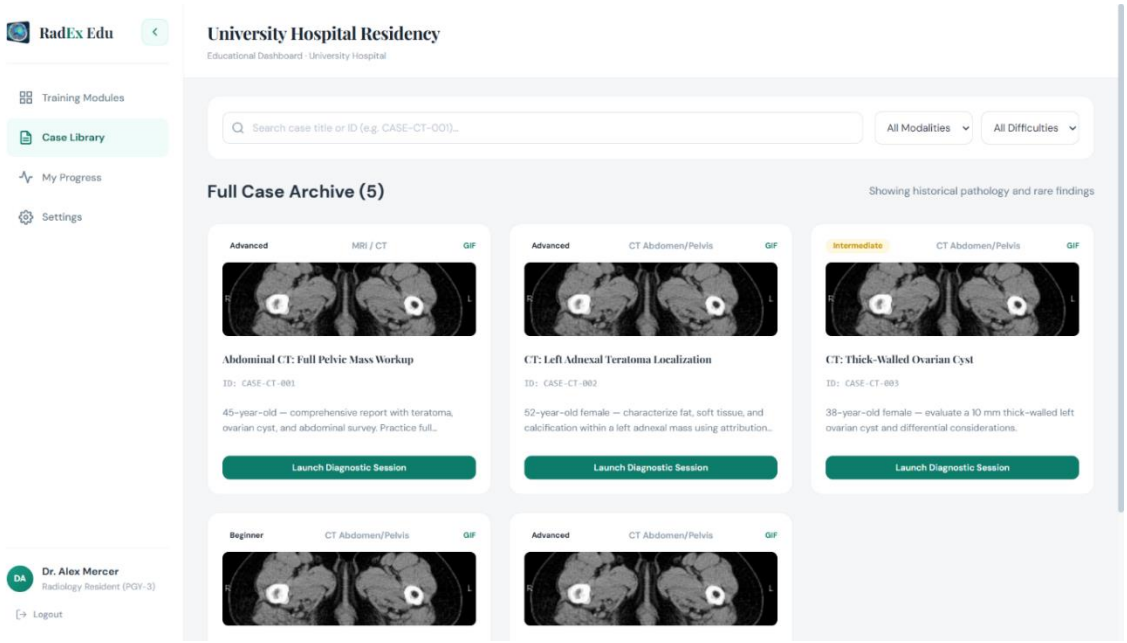
1-لوحة التدريب والتعليم الطبي

تعرض تقدم الطلبة والمقيمين داخل المسارات التعليمية والتدريبية.



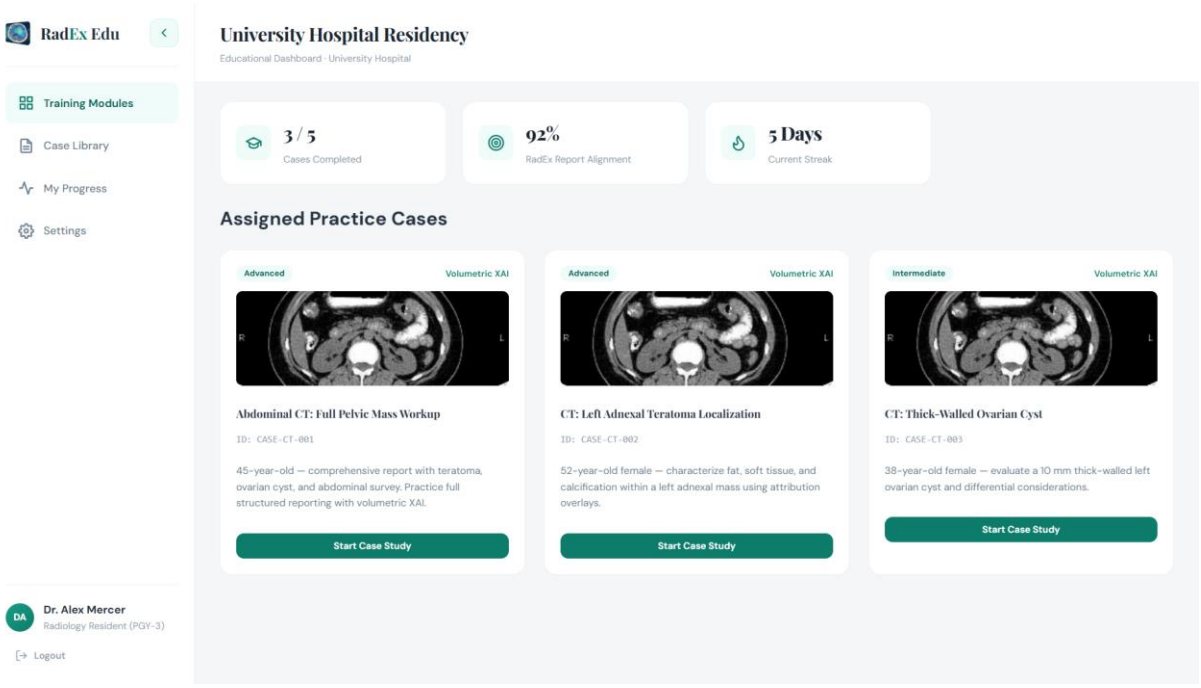
2-استعراض الحالات الطبية التعليمية

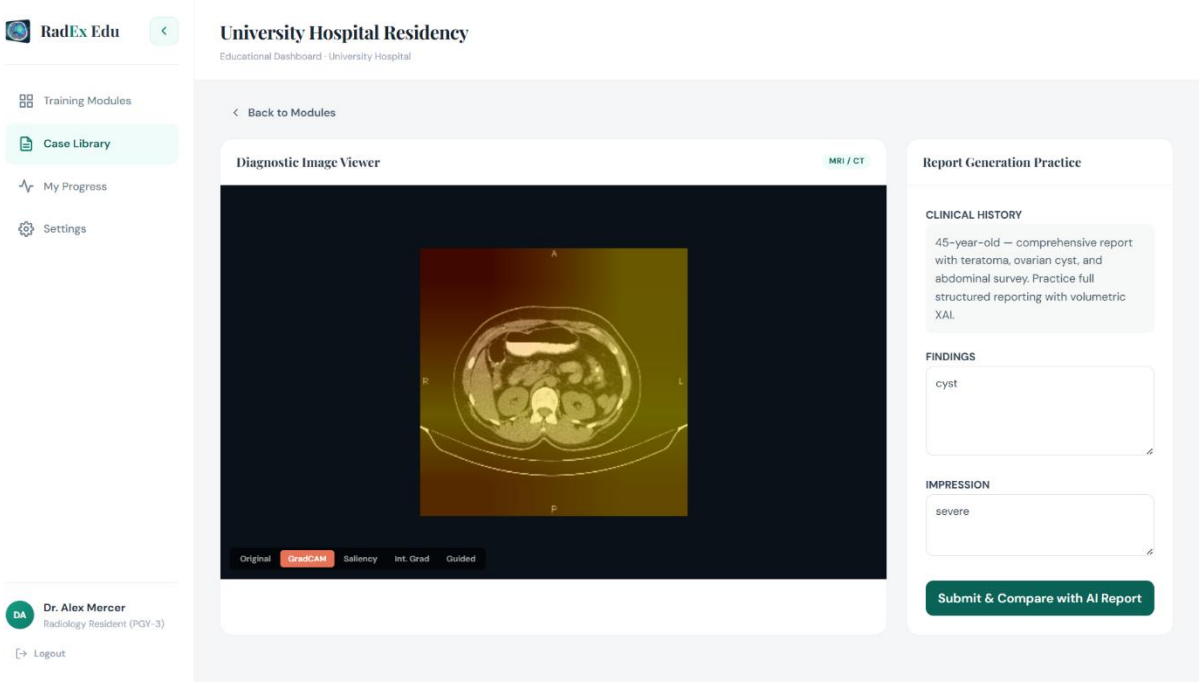
توفر مكتبة لحالات سريرية متنوعة يمكن استخدامها في التدريب والتقييم الأكاديمي.



3-جلسات تحليل الصور التعليمية

تمكّن الطلبة من تحليل الصور الطبية وكتابة تقاريرهم الخاصة قبل مقارنتها بنتائج الذكاء الاصطناعي.





تعكس الواجهات المعروضة مدى تكامل منصة **RadEx** في توفير بيئة رقمية ذكية تجمع بين تحليل الصور الطبية، التفسير البصري لقرارات الذكاء الاصطناعي، والتوليد الآلي للتقارير الطبية ضمن واجهات حديثة وسهلة الاستخدام.

الملحق رقم 01

ميزانية المؤسسة الناشئة

ميزانية المؤسسة الناشئة ACTIF

بالآلاف دج	N-2	N-1	N	N+1	N+2	N+3	N+4	N+5
الأصول غير الملموسة	-	-	220	180	150	120	100	80
الأصول المادية	-	-	550	590	570	500	370	190
الأراضي	-	-	-	-	-	-	-	-
المباني	-	-	-	-	-	-	-	-
تجهيزات وأصول مادية أخرى	-	-	550	590	570	500	370	190
أصول الامتياز والاستغلال	-	-	-	-	-	-	-	-
الأصول قيد الإنجاز	-	-	-	-	-	-	-	-
الأصول المالية	-	-	-	-	-	-	-	-
سندات المساهمة	-	-	-	-	-	-	-	-
مساهمات وحقوق مرتبطة	-	-	-	-	-	-	-	-
سندات مالية أخرى	-	-	-	-	-	-	-	-
قروض وأصول مالية طويلة الأجل	-	-	-	-	-	-	-	-
ضرائب مؤجلة أصول	-	-	-	-	-	-	-	-
إجمالي الأصول غير الجارية	**-**	**-**	**770**	**770**	**720**	**620**	**470**	**270**
المخزونات	-	-	-	-	-	-	-	-
الذمم المدينة والأصول المشابهة	-	-	215	732	1,489	2,477	3,644	4,986
العملاء	-	-	215	732	1,489	2,477	3,644	4,986
مدينون آخرون	-	-	-	-	-	-	-	-
الضرائب والرسوم	-	-	-	-	-	-	-	-
ذمم وأصول متداولة أخرى	-	-	-	-	-	-	-	-

بالآلاف دج	N-2	N-1	N	N+1	N+2	N+3	N+4	N+5
النقدية وما يعادلها	-	-	308	1,640	6,052	13,382	20,519	31,624
استثمارات وأصول مالية قصيرة الأجل	-	-	-	-	-	-	-	-
الخزينة	-	-	308	1,640	6,052	13,382	20,519	31,624
إجمالي الأصول الجارية	**-**	**-**	**523**	**2,372**	**7,541**	**15,859**	**24,163**	**36,610**
إجمالي الأصول	**-**	**-**	**1,293**	**3,142**	**8,261**	**16,479**	**24,633**	**36,880**

ميزانية المؤسسة الناشئة PASSIF

بالآلاف دج	N-2	N-1	N	N+1	N+2	N+3	N+4	N+5
رأس المال وحقوق الملكية								
رأس المال المصدر	-	-	500	500	500	500	500	500
رأس المال غير المطلوب	-	-	-	-	-	-	-	-
فروقات إعادة التقييم	-	-	-	-	-	-	-	-
الاحتياطات والفوائض	-	-	500	693	2,492	5,511	11,579	18,533
النتيجة الصافية للسنة	-	-	193	1,799	5,019	10,068	11,954	16,947
الأرباح المرحلة	-	-	-	-	-	-	-	-
إجمالي حقوق الملكية	**-**	**-**	**1,193**	**2,992**	**8,011**	**16,079**	**24,033**	**35,980**
الخصوم غير الجارية								
القروض والديون المالية	-	-	-	-	-	-	-	-
الضرائب المؤجلة خصوم	-	-	-	-	-	-	-	-
ديون غير جارية أخرى	-	-	-	-	-	-	-	-
المؤونات والمنتجات المسجلة مسبقاً	-	-	-	-	-	-	-	-
إجمالي الخصوم غير الجارية	**-**	**-**	**0**	**0**	**0**	**0**	**0**	**0**
الخصوم الجارية								
الموردون والحسابات المرتبطة	-	-	60	100	200	350	550	800
الضرائب	-	-	0	0	0	0	0	0
ديون أخرى	-	-	40	50	50	50	50	100
خزينة الخصوم	-	-	-	-	-	-	-	-
إجمالي الخصوم الجارية	**-**	**-**	**100**	**150**	**250**	**400**	**600**	**900**
إجمالي الخصوم	**-**	**-**	**1,293**	**3,142**	**8,261**	**16,479**	**24,633**	**36,880**
التحقق من توازن الأصول والخصوم	**-**	**-**	**متوازن**	**متوازن**	**متوازن**	**متوازن**	**متوازن**	**متوازن**

الملحق رقم 02

حسابات النتائج المتوقعة

بالآلاف دج	N-2	N-1	N	N+1	N+2	N+3	N+4	N+5
مبيعات والمنتجات الملحقة	-	-	1,981	6,754	13,746	22,860	33,636	46,020
تغير مخزون المنتجات قيد الإنجاز	-	-	-	-	-	-	-	-
الإنتاج المثبت	-	-	-	-	-	-	-	-
إعانات الاستغلال	-	-	0	0	0	0	0	0
إنتاج الدورة	**-**	**-**	**1,981**	**6,754**	**13,746**	**22,860**	**33,636**	**46,020**
المشتريات المستهلكة (COGS %20)	-	-	396	1,351	2,749	4,572	6,727	9,204
الخدمات الخارجية والاستهلاكات الأخرى	-	-	1,291	2,254	3,408	4,400	5,900	7,900
استهلاك الدورة	**-**	**-**	**1,687**	**3,605**	**6,157**	**8,972**	**12,627**	**17,104**
القيمة المضافة للاستغلال	**-**	**-**	**294**	**3,149**	**7,589**	**13,888**	**21,009**	**28,916**
أعباء المستخدمين والرواتب	-	-	0	1,200	2,400	3,600	4,800	6,000
الضرائب والرسوم المماثلة	-	-	0	0	0	0	0	0
فائض الاستغلال الخام (EBE)	**-**	**-**	**294**	**1,949**	**5,189**	**10,288**	**16,209**	**22,916**
المنتجات التشغيلية الأخرى	-	-	0	0	0	0	0	0
الأعباء التشغيلية الأخرى	-	-	0	0	0	0	0	0
مخصصات الاهتلاك والمؤونات	-	-	100	120	150	200	250	300
استرجاع خسائر القيمة والمؤونات	-	-	-	-	-	-	-	-
النتيجة التشغيلية	**-**	**-**	**193**	**1,829**	**5,039**	**10,088**	**15,959**	**22,616**
المنتجات المالية	-	-	0	0	0	0	0	0
الأعباء المالية	-	-	0	30	20	20	20	20
النتيجة المالية	**-**	**-**	**0**	**-30**	**-20**	**-20**	**-20**	**-20**

بالآلاف دج	N-2	N-1	N	N+1	N+2	N+3	N+4	N+5
النتيجة العادية قبل الضريبة	**-**	**-**	**193**	**1,799**	**5,019**	**10,068**	**15,939**	**22,596**
الضريبة المستحقة على النتيجة العادية	-	-	0	0	0	0	3,985	5,649
الضريبة المؤجلة	-	-	-	-	-	-	-	-
النتيجة الصافية للأنشطة العادية	**-**	**-**	**193**	**1,799**	**5,019**	**10,068**	**11,954**	**16,947**
المنتجات الاستثنائية	-	-	-	-	-	-	-	-
الأعباء الاستثنائية	-	-	-	-	-	-	-	-
النتيجة الاستثنائية	**-**	**-**	**-**	**-**	**-**	**-**	**-**	**-**
النتيجة الصافية للسنة	**-**	**-**	**193**	**1,799**	**5,019**	**10,068**	**11,954**	**16,947**

الملحق رقم 03

حسابات الخزينة

البنود	N-2	N-1	N	N+1	N+2	N+3	N+4	N+5
التدفقات النقدية الناتجة عن الأنشطة التشغيلية								
النتيجة الصافية للسنة	-	-	193,000	1,799,000	5,019,000	10,068,000	11,954,000	16,947,000
الاستهلاكات والمؤونات	-	-	100,000	120,000	150,000	200,000	250,000	300,000
تغير الضرائب المؤجلة	-	-	0	0	0	0	0	0
تغير المخزون	-	-	0	0	0	0	0	0
تغير حسابات الزبائن والذمم المدينة	-	-	-215,000	-517,000	-757,000	-988,000	-1,167,000	-1,342,000
تغير الموردين والديون الأخرى	-	-	100,000	50,000	100,000	150,000	200,000	300,000
فائض أو عجز التنازل الصافي من الضرائب	-	-	0	0	0	0	0	0
التدفقات التشغيلية (A)	**0**	**0**	**178,000**	**1,452,000**	**4,512,000**	**9,430,000**	**11,237,000**	**16,205,000**

البنود	N-2	N-1	N	N+1	N+2	N+3	N+4	N+5
التدفقات النقدية الناتجة عن الاستثمار								
اقتناء التجهيزات والأصول	-	-	-870,000	-120,000	-100,000	-100,000	-100,000	-100,000
مداخيل التنازل عن الأصول	-	-	0	0	0	0	0	0
أثر تغيرات محيط الاستثمار	-	-	0	0	0	0	0	0
التدفقات الاستثمارية (B)	**0**	**0**	**-870,000**	**-120,000**	**-100,000**	**-100,000**	**-100,000**	**-100,000**
التدفقات النقدية الناتجة عن التمويل								
توزيعات الأرباح	-	-	0	0	0	-2,000,000	-4,000,000	-5,000,000
دعم وتمويل أولي	-	-	500,000	0	0	0	0	0
مساهمة المؤسسين	-	-	500,000	0	0	0	0	0
تمويل الحساب الجاري للشركاء	-	-	0	0	0	0	0	0
تسديدات رأس المال	-	-	0	0	0	0	0	0
تسديد الحساب الجاري للشركاء	-	-	0	0	0	0	0	0
التدفقات التمويلية (C)	**0**	**0**	**1,000,000**	**0**	**0**	**-2,000,000**	**-4,000,000**	**-5,000,000**
التغير في الخزينة خلال الفترة (A+B+C)	**0**	**0**	**308,000**	**1,332,000**	**4,412,000**	**7,330,000**	**7,137,000**	**11,105,000**
خزينة بداية الفترة	-	-	0	308,000	1,640,000	6,052,000	13,382,000	20,519,000
خزينة نهاية الفترة	-	-	308,000	1,640,000	6,052,000	13,382,000	20,519,000	31,624,000
التغير في الخزينة	**0**	**0**	**308,000**	**1,332,000**	**4,412,000**	**7,330,000**	**7,137,000**	**11,105,000**

الملحق رقم 04

نموذج العمل التجاري

شرائح العملاء
- العيادات الخاصة
- المستشفيات العمومية EPH / EHS
- المستشفيات الجامعية CHU
- كليات الطب والمراكز الجامعية
- المخابر البحثية الطبية

العلاقات مع العملاء
- دعم تقني مستمر
- تكوين ومرافقة أثناء الاستخدام
- تخصيص بعض الخصائص حسب المؤسسة
- جمع الملاحظات لتحسين المنصة
- تجربة استخدام بسيطة وسريعة

القنوات
- البيع المباشر B2B
- المؤتمرات والفعاليات الطبية
- الشراكات مع المستشفيات الجامعية
- LinkedIn والتواصل المهني
- التعاون مع الأطباء والخبراء

القيمة المقترحة
- تشخيص أسرع وأكثر دقة للأشعة الطبية
- ذكاء اصطناعي تفسيري قابل للشرح (XAI)
- احترام سيادة البيانات الصحية الجزائرية
- تكلفة أقل مقارنة بالحلول الأجنبية
- واجهات مخصصة حسب نوع المؤسسة

الأنشطة الرئيسية
- تطوير وصيانة محرك XAI
- معالجة وتحليل الصور الطبية
- تشغيل وصيانة المنصة
- توليد التقارير الطبية
- دعم العملاء وتحسين الخصائص

الموارد الرئيسية
- فريق تقني متخصص
- محرك ذكاء اصطناعي مطور داخلياً
- بنية سحابية مؤهلة للبيانات الصحية
- منصة ويب طبية
- خبرة أكاديمية وبحثية

الشركاء الرئيسيون
- مزودو الخدمات السحابية
- المستشفيات الجامعية CHU
- Algeria Venture
- ANADE و Startup Label
- شركاء أكاديميون وطبيون

مصادر الإيرادات
- اشتراكات SaaS شهرية وسنوية
- عقود مؤسساتية مخصصة
- خدمات تكامل PACS / API
- خدمات التكوين والدعم
- برنامج الشركاء التأسيسيين

هيكل التكاليف
- البنية التحتية السحابية
- التسويق والمبيعات B2B
- المصاريف الإدارية والقانونية
- الاستهلاكات التقنية